DOCUMENT RESUME

ED 195 089 EC 130 866

AUTHOR May, Marcia J., Ed.
TITLE Evaluating Handicapped Children's Early Education Programs. Proceedings of the WESTAR Evaluation Workshop (Albuquerque, New Mexico, January 31-February 2, 1979).
INSTITUTION Washington Univ., Seattle. Western States Technical Assistance Resource.
SPONS AGENCY Office of Special Education (ED), Washington, D.C.
PUB DATE 80
CONTRACT 300-77-0508
NOTE 125p.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Disabilities; Early Childhood Education; *Evaluation Methods; Models; *Parent Education; Planning; *Program Evaluation; *Staff Development
IDENTIFIERS *Handicapped Childrens Early Education Program

ABSTRACT

Seven papers from a 1979 workshop address program evaluation factors (specifically planning) in the Handicapped Children's Early Education Program. R. Sheehan ("Evaluation Strategies in Early Childhood Education: A National Perspective") summarizes the state of the art and lists seven guidelines, such as that evaluation efforts must reflect qualitative changes in children's performance. In "Practical Program Evaluation: Many Problems and a Few Solutions," O. White details the advantages of curriculum referenced approaches to evaluation, while R. Sheehan ("Measuring Child Progress: Large Group Design and Norm Referenced Alternatives") reviews other evaluation designs. M. Stevens and R. Kroth address problems and issues in "Evaluating Parent Training Programs: Some Considerations." Inservice training is the focus of "Evaluating Staff Development" by I. Lynch and V. Lynch. The final paper ("Evaluating the Impact of Demonstration and Dissemination Activities" by P. Bradley) considers components of impact evaluation studies and describes conditions which should be met before undertaking one. (CL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

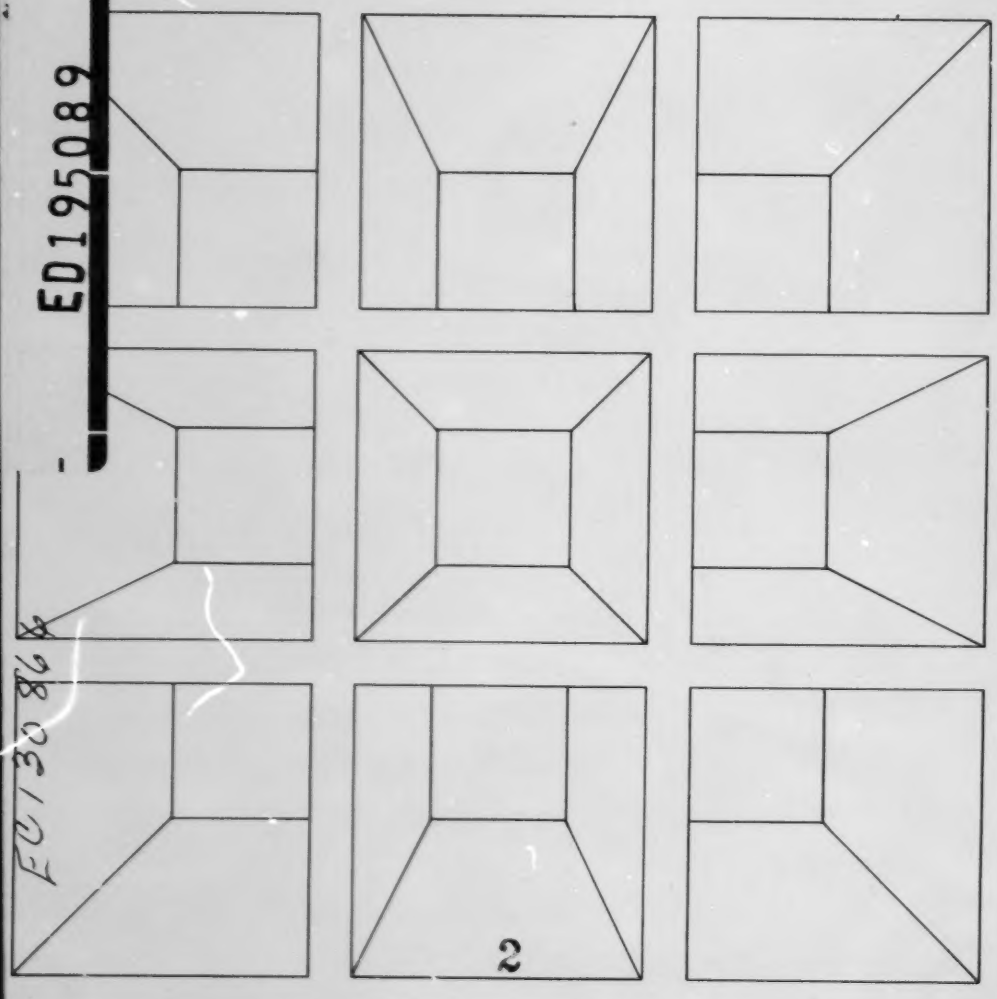
Evaluating Handicapped Children's Early Education Programs

Proceedings of
the 1979 WESTAR
Evaluation Workshop
Jan 31-Feb 2, 1979
Albuquerque, New Mexico

Edited by Marcia J. May
Produced by
the Western States
Technical Assistance
Resource (WESTAR)

ED195089

EC130868



Word Processing by Cheryl Yates

This document was produced under contract number 300-77-0508 from the Office of Special Education, U.S. Department of Education. The opinions expressed herein do not necessarily reflect the position or policy of the U.S. Department of Education, and no official endorsement by them should be inferred.

Printed in the United States of America

1980

Contents

v Contributors

vii Foreword, William W. Swan

1 Introduction: The First Steps in Evaluation
Joan M. A. Karp

5 Evaluation Strategies in Early Childhood Education: A National Perspective
Robert Sheehan

21 Practical Program Evaluation: Many Problems and a Few Solutions
Owen R. White

55 Measuring Child Progress: Large Group Design and Norm-Referenced Alternatives
Robert Sheehan

79 Evaluating Parent Training Programs: Some Considerations
Meave V. Stevens and Roger Kroth

93 Evaluating Staff Development
Linda L. Lynch and Valerie Lynch

113 Evaluating the Impact of Demonstration and Dissemination Activities
Paul A. Bradley

4 JUL 21 1980

Contributors

ROBERT SHEEHAN
Assistant Professor
Dept. of Child Development
and Family Studies
Purdue University
West Lafayette, Indiana 47907

OWEN R. WHITE
Coordinator of Planning,
Development and Evaluation
Experimental Education Unit
Child Development and Mental
Retardation Center
University of Washington
Seattle, Washington 98195

ROGER KROTH
Associate Professor
Dept. of Special Education
Director, Parent Involvement
Center
University of New Mexico
1700 Pennsylvania N.E.
Albuquerque, New Mexico 87110

MEAVE V. STEVENS
Research Assistant
Dept. of Special Education
University of New Mexico
Albuquerque, New Mexico 87131

LINDA L. LYNCH
Supervisor
Division of Special Services
Superintendent of Public
Instruction
Olympia, Washington 98504

VALERIE LYNCH
Project Manager
Center for the Severely
Handicapped, JD-05
University of Washington
Seattle, Washington 98105

PAUL A. BRADLEY
Associate Professor
Coordinator of Research and
Evaluation Methodology
School of Education
University of Colorado
at Denver
1100 14th Street
Denver, Colorado 80202

Foreword

Program evaluation has been an integral and critical component of all Handicapped Children's Early Education Program (HCEEP) projects since the initiation of the program in 1969. The directors of those first demonstration projects still remember the emphasis placed on program evaluation at the first Project Directors' Meeting in 1969. That emphasis has been maintained through the past ten years and program evaluation will continue to be an extremely important component of all HCEEP projects.

Today each project director is confronted with essentially the same dilemma as in the past: "How can I most effectively meet the needs of all my project's target audiences in order to develop my project model, to demonstrate and disseminate my project model, to obtain continuation funding for my project's direct services, and to stimulate others to use my model to serve additional preschool handicapped children and their families--all on limited project resources?" Solutions to this dilemma require knowledge of one's target audiences and an anticipation of their informational needs and criteria for success. Today, however, project directors must also design program evaluations of project components to meet the requirements of increasingly sophisticated target audiences. HCEEP has matured over the past ten years in terms of services to children, curriculum, services to parents, staff development, demonstration/dissemination and the many other project components; however, target audiences have increased not only their awareness, knowledge, and skills in the areas of early childhood education but also their critical analysis of project's reported effectiveness with handicapped children and

their families. The project director and the project staff must analyze needs and allocate limited program evaluation resources to meet these needs.

The articles by the contributors to this publication have not only chronicled a perspective of the ten-year history of program evaluation in HCEEP projects but have also suggested directions for future efforts. The reviews and discussions of alternative statistical designs, including practical examples and references, are most valuable for project directors and their staffs. Two issues of continuing concern are the measurement of child change while accounting for the maturational impact on that change and the comparison of one treatment approach against another. This latter issue will perhaps become increasingly significant as more mandates for preschool education for the handicapped are enacted and such mandates become commonly accepted and supported. The information provided on parent programs, staff development and demonstration/dissemination are also significant in guiding evaluative efforts in these areas.

This document should be of significant assistance to project directors and their staffs as they continue to implement program evaluations to assist in developing the most effective models for serving preschool handicapped children and their families.

William W. Swan

Acting Chief
Program Development
Branch
Division of Innovation
and Development
Office of Special Education

Introduction: The First Steps in Evaluation

Joan M. A. Karp

An important step in the development of model services for handicapped preschool children and their families is the documentation of the effectiveness of those services. Toward this end, WESTAR sponsored a workshop on evaluation for its demonstration project clients in the Handicapped Children's Early Education Program (HCEEP), on January 31-February 2, 1979 in Albuquerque, New Mexico. The programs within HCEEP are funded to develop and provide model services to handicapped children and their families while also developing the capability for staff training, demonstration and dissemination. Since these programs are multifaceted, the workshop focused on measurement and evaluation designs for demonstrating child progress and determining the effectiveness of other program components. It also provided participants with an opportunity to discuss the plans for their project's evaluation with an evaluation specialist.

The presentations made at the workshop are compiled in this document and specifically address the first phase of program evaluation, that of planning. They describe the types of evaluation designs most commonly used by handicapped children's early education projects and the considerations which must be made in choosing among them. The papers do not discuss how to collect, analyze, or report data (subsequent phases in evaluation), but rather how to conceptualize the overall evaluation design necessary for the population of children being served and the component or components being examined.

Joan M. A. Karp is Acting Director of WESTAR.

Choosing an evaluation design is only one part of the planning phase. The project must first specify the purpose and target audience of the evaluation in order to determine the type of evaluation necessary. The type of effectiveness data required to convince a funding agency of continued support is usually very different from that required by a scholarly journal.

The next step is to identify the program's objectives. The program objectives determine what, in fact, will be evaluated. A program whose main objective is to increase the communication skills of hearing impaired preschoolers will obviously focus its evaluation efforts in this area rather than some other area such as the child's self-help skills. Treatment or independent variables must then be described in detail so that they can be measured appropriately and ultimately replicated. In the previous example, the program goal is accomplished by training the child's parents. The scope, sequence and strategies taught to the parents must be carefully described so that other teachers and researchers using the program could achieve the same results. The selection of the dependent variable is the next important step. The dependent variable is the presumed effect of the treatment. In this case there are four possible areas affected by the program: the children's hearing aid management, auditory awareness, language skills, and preacademic cognitive skills. For each of these possible effects an instrument would be selected and a measurement made.

The next step in the planning phase is to determine the type of evaluation design which will answer the questions about effectiveness. The papers by Rob Sheehan and Owen White present two different approaches to design, group designs and individual or single subject designs, respectively. Both approaches have been used successfully with preschoolers, but as Sheehan and White indicate, each approach has limitations as well. Stevens and Kroth outline a variety of possible approaches to evaluating parent programs resulting from their work at the Parent Involvement Center. They describe a range of possible avenues for evaluating parent programs without using standardized measures. Lynch and Lynch outline a staff development process which builds evaluation into its format. Bradley raises a number of issues in carrying out an impact evaluation for dissemination activities.

The final steps in the planning phase are to assign responsibilities and timelines to each step in the process

described. If HCEEP projects complete an entire evaluation plan by the end of their first year of funding, then they are well on their way towards determining the effectiveness of their programs.

As indicated earlier, the papers presented at this workshop were designed for a very specific target audience and as such are not definitive works on evaluation strategies. The contributors have presented documentation of new evaluation approaches in the field of special education which are continually being refined and elaborated. Hopefully these proceedings will be helpful as a resource to projects for planning their evaluations.

Evaluation Strategies in Early Childhood Education: A National Perspective

Robert Sheehan

Good evening. I add my welcome to this topical conference on evaluation. The area of evaluation is probably one of the most controversial in the field of education today. I hope that as this conference proceeds we will all learn more about it. As a cautious introduction however, I would like to share with you a saying that I saw posted on a door at Georgia State University in Atlanta: "Learning is the dawning recognition of how little one knows about something." I hope that our learning efforts are not quite as pessimistic as this, but the caution is quite appropriate that the subject may get much more difficult before it gets any easier.

Let me first introduce myself and share with you how I come to be addressing you this evening. Following this I will suggest some definitions that we will be using during this conference and will describe the state of the art of evaluation of early childhood special education programs. Finally, I will discuss three major themes which I am learning more about as I move about the field of early childhood special education.

For the past several years I have been serving as an evaluation consultant to projects such as yours around the country. My task and that of several other persons, such as Owen White who is here with us now, has been to assist projects in the development and implementation of evaluation plans. For the past year, I have been emphasizing a different approach to this task by working on the development of state-of-the-art papers and basic research on evaluation in early childhood special education with Project REACH at UCLA, which is one of the four

early childhood research institutes funded by the Bureau of Education for the Handicapped. My work at UCLA has brought about a different perspective and is allowing me to ask some of the "why" questions with regard to evaluation methodology that now seem so timely.

DEFINITION OF TERMS

Let me try to set the tone for this conference by suggesting three definitions. The purpose of these definitions is for us to be open-minded about evaluation in early childhood special education and to establish a common ground. The terms we will define are "evaluation," "instrument," and "early childhood special educator."

Any review of the literature on evaluation reveals that a considerable debate exists about the difference (or non-difference) between evaluation and research. In a very controversial article, Kerlinger (1977) argues that basic research is, in the long run, more productive than applied research such as evaluation. An interesting counterpoint is offered by Mitroff and Bonoma (1978), who argue that in fact the processes which constitute evaluation and laboratory research are so similar that the distinction between them may not be relevant. This counterpoint is the one that I prefer. I would argue that evaluation (and all research) is a process by which data are collected and used for decision making. This decision making may occur in the laboratory or in the field. Ultimately, what you in early childhood special education are doing with evaluation is planning for decision making.

There is also a debate among professionals about what constitutes an instrument. No one would deny that the Stanford Binet Intelligence Scale is an instrument, but would you call a checklist or a rating scale an instrument? I would say yes. I suggest that an instrument is a data collection guide which insures that information yielded by one person is qualitatively similar to the information provided by another person using the same guide on the same child. This working definition suggests then that the term instrument can be used for rating scales, developmental checklists, questionnaires, tests, and much more. As we go through the rest of this conference, we repeatedly will be coming back to the

question of instrumentation, and I hope that we can keep this broad view in mind.

This leaves me with the final definition, that of an early childhood special educator. I would like to suggest a definition which emphasizes the role that you play in the evaluation of your programs. An early childhood special educator is a decision maker in the field of education. As you move into the sessions of this conference, please hold your head high and remember this definition.

There are implications to the definition of an early childhood special educator as a decision maker. The first of these is that you must learn what questions to ask in order to make decisions. The second is that you must know when to turn to an expert and when to work decisions out for yourself. Let me elaborate on this a bit. As evaluation consultants move from one project to the next, they are continually asked, "What design should I use?" or "Which instrument is best?" There are no general answers to these questions, and the sooner that we all realize that, the better. On the other hand, if you were to ask, "What do I need to know to select an instrument?", answers would come more quickly. You are the decision makers. As evaluators we will help you define your questions and direct your search rather than allow you to walk away from this conference with answers which may or may not apply to your situation. As in any field, decision makers need methods and procedures to help make decisions rather than specific answers to questions which seem apparent now but are of limited use in the future.

WHY EVALUATE EARLY CHILDHOOD SPECIAL EDUCATION?

The question of whether or not we engage in evaluation efforts in early childhood special education seems to be a decided issue. All First Chance projects must respond to the concerns of the Bureau of Education for the Handicapped for program evaluation. This concern applies specifically to the assessment of children's progress (Sec. 3.10 in RFP CFDA #13.44A); however, other aspects of the program, such as parent-family participation, must also be evaluated.

To put it very pragmatically, you must evaluate because you have proposed to your funding agency that you will evaluate. A second pragmatic view is that your presence at this conference indicates a willingness to engage in program evaluation. However, there are some other very good reasons why you as decision makers should evaluate. Perhaps the best of these is that you need data to adequately make your decisions. We must move carefully however, because there are some very strong reasons why certain types of evaluation efforts may not be appropriate. These reasons, in almost all cases, reflect our lack of knowledge about how children, and handicapped children in particular, develop over time. In addition, we want to avoid collecting data only for the sake of collecting data. In introducing this conference I want to support the overall importance of evaluation (and I think you would agree), but I also want us to be cautioned about moving too quickly in an area that we are just beginning to understand.

THE STATE OF THE ART

The best way that I have found to view the field of early childhood special education is through a matrix common to most of you in the First Chance network. This matrix separates the field into four types of service. They include service to children, service to parents and families, staff development, and outreach/dissemination activities. My impression is that the four areas, as I have mentioned them, are rank-ordered in the emphasis usually given to them in evaluation efforts. Let me clarify this. I don't think that evaluation of staff development is necessarily less important than evaluation of the services to children component, but it does seem to receive less attention by the field.

SERVICES TO CHILDREN

There are several interesting developments in the documentation of handicapped children's progress. I will be sharing some of these with you in my address tomorrow, but let me call your attention to others as an introduction. The first of these is the successful submission

to the Joint Dissemination Review Panel by Teaching Research of Monmouth, Oregon of an evaluation design utilizing individual time-series data. Those of you who belong to the Council for Exceptional Children, Division for Early Childhood, will find a write-up of these results in the Communicator (1978, 5(1), 15-16). If the questions about the children you work with can be addressed by such a design, it is an exciting alternative to the more traditional group-style evaluation designs.

Another noteworthy development in the area of documentation of child progress is the increasing use of prediction-type evaluation designs and instruments. The Battelle national evaluation of the First Chance network (Stock, Wnek, Newborg, Schenck, Gabel, Spurgeon, & Ray, 1976) and several other successful Joint Dissemination Review Panel submissions have utilized these designs. As the number of children who might serve as research controls decreases (since P.L. 94-142 mandates services for all children), researchers are turning more and more to comparisons of children's progress with a statistical or prediction control group. I will discuss the strengths and drawbacks of these approaches more tomorrow, but the fact that these designs are being used should be noted now.

A third development which I find quite heartening is that the number of instruments available to early childhood special educators seems to be increasing to such a point that projects are able to turn to them rather than to develop their own project-specific instruments. At UCLA we conducted a review of infant assessment instruments and found over 300 of them at first count. The psychometric properties of these instruments vary widely from full, adequate psychometric data to no psychometric data at all. Their existence, however, provides us with a pool from which to draw better instruments to suit our purposes.

A fourth development, which is notable by its absence, is that there is still little emphasis in the field to the documentation of the specific intervention being provided to children by model programs. The word "model" usually connotes the existence of a set of parameters which describe very specifically what goes on during the course of the day, week, and year (McDaniels, 1977). This might be called the surface structure of the model. "Model" also usually connotes a theoretical rationale which supports those parameters

(Kopp & Yoshioka-Maxwell, 1978). This might be called the deep structure. As I survey the efforts of model programs in the First Chance network to document their services to children, the absence of these parameters and theoretical foundations is a cause for concern.

SERVICES TO PARENTS AND FAMILIES

The evaluation efforts of projects in the area of services to parents and families vary widely, depending upon the type of program provided for them. In many cases projects seem content to document the satisfaction of parents and families with the services that are being delivered to children. This is actually a two-part issue. Many projects are not providing any direct services to parents and families. In addition, fewer programs are evaluating the parent and family component beyond this simple documentation of satisfaction.

There is, however, a strong rumbling in the field of early childhood special education that as the children with whom we are work get younger, the direct intervention efforts move from the child (or infant) to the mother (or primary caretaker). This then requires that our evaluation efforts focus on change in parents and families as well as change in the interactional patterns which exist between children and those persons caring for them (Kopp, 1978).

The field of early childhood special education has been aware of the ecological nature of our intervention (Bronfenbrenner, 1974). It is clear that you as projects are responding to this concern for ecological intervention. What is less clear is if our evaluation capabilities can presently document this type of impact.

I ask you to consider the research of Lazar, who with the assistance of several colleagues, is following up on the children who participated in pioneering early intervention studies in the late 1950s and early '60s (Lazar, Hubbell, Murray, Rosche, & Royce, 1977). Significant differences have emerged in the attitudes of their parents toward education and the role that children play in the household (for example, T.V. watching).

I think that it is time for early childhood special educators to look to supporting fields such as social work and public health for assistance in designing our evaluation efforts. This assistance would be particularly useful

in the design of instruments. We should also continue our efforts to document the changing nature of handicapped children's interactions with their families.

STAFF DEVELOPMENT

The staff development component of First Chance projects may be the key to our continuing success in early childhood special education. Evaluators seem to place minimal emphasis on the evaluation of this component, and I hope that a conference such as this can help to reverse this trend.

In considering the work of other intervention efforts such as Head Start and Follow Through, I would like to point out that, in most cases, the dramatic impact of a model program is to awaken and energize the staff of that program. Ultimately, this learning process on the part of the staff must have a positive impact on our children's growth and development.

I would like to see us move far beyond the simple documentation that staff development activities have occurred and consider the possibility of documenting the very real changes that occur in the professional abilities of our staffs as they participate in a First Chance project. Consider the fact that most projects begin with a totally new program staff at the beginning of their three-year funding period. Staff come into programs with both strengths and weaknesses. Let's work to document incoming strengths and use that information to ameliorate the weaknesses and evaluate the success of our efforts.

One activity that I have found quite useful is to periodically ask staff to identify their own and others' strengths and weaknesses and then to chart the changing nature of these responses over the months of a program's existence. Business and industry have been using this method for years, and I think we can learn from their efforts.

OUTREACH/DISSEMINATION

The area of outreach/dissemination seems to receive the least amount of attention and yet it is an area that has the most potential for documenting change. I say this because most projects do not begin to carry out

these activities until midway through their second year and don't really get going until their third year. It appears to me that this is an excellent opportunity to show change before and after the activities begin!

Once First Chance projects begin to realize that outreach/dissemination activities can be evaluated, they will identify their potential audience long before they begin serving them. They document the needs of that audience, and as they begin to serve them, continue their documentation efforts to identify the meeting of needs.

If you go back and read some of the original discussion surrounding the passage of the Handicapped Children's Early Education Assistance Act of 1969, you will find very heavy emphasis on the demonstration capabilities of this program. As I consider other federally-funded programs, the First Chance projects seem unique in their legislative mandate to engage in outreach/dissemination activities. I would hope that conferences such as this would reinforce our willingness to evaluate these activities.

LESSONS BEING LEARNED

As I move about the field of early childhood special education and consider the current state of the art in evaluation efforts, I find myself returning time and time again to some common themes. These themes have to do with the blending of research design concepts, statistical concepts, and the purposes of evaluation in early childhood special education. I would like to share with you a synthesis of my thinking about these themes with the hope that they will influence your thinking throughout the rest of this conference and when you return to your programs.

To help communicate these themes, I have clustered them into three major categories and put them into the form of "lessons I am learning." If you recall my opening remarks, they are not necessarily truths. Rather, they are questions over which I puzzle and solutions at which I have tentatively (and perhaps temporarily) arrived. The three clusters include themes dealing with independent variables, dependent variables, and the meaning of significance in the field of early childhood special education.

A FOCUS ON INDEPENDENT VARIABLES

Lesson #1
Noncategorical education should not mean non-categorical evaluation.

The first lesson which I am learning pertains to the noncategorical nature of First Chance projects. A review of their abstracts reveals that very few projects are categorical in nature. This makes a lot of sense as more and more programs for handicapped children are funded. With more children being served, a program which seeks full enrollment must be broadly based. In addition, the emphasis in the field on a least restrictive environment dictates a fair amount of heterogeneity of our children.

The lesson which I am learning is that noncategorical education does not and should not mean noncategorical evaluation. Gather as much demographic information about your children as you can. The more information you have to describe children, the better. In research design terms, what I am suggesting is that you reduce the unexplained variation in your children's performance by including information pertaining to children's handicapping conditions or performance as independent or control variables in your evaluation designs.

This lesson suggests that if you are serving children with visual impairment, speech delay, and orthopedic complications (or any other identifiable condition), you may choose a design which groups the children together or, at the very least, highlights the performance of any one particular subgroup of children. Let me stress that this grouping is only of children's scores and is only for data analysis and report-writing purposes. Your program can be noncategorical in educational approach and still use these types of designs.

Lesson #2
Children receiving your intervention may be better off than children receiving an alternate form of intervention. An evaluation effort should make this comparison.

A second lesson which I am learning also pertains to the selection of independent variables. The traditional research design calls for comparison of the performance of children who are receiving intervention with children who are receiving no intervention. My review of the current activities in the field have failed to uncover any projects which have randomly assigned children to such situations and only one project that has made use of this type of design with intact groups. With the increasing implementation of P.L. 94-142 and various state mandates, there will be no possibility of such comparisons, as all children will be served. Rather than immediately turn to research designs which do not require comparison groups, I would remind you that as model projects you hopefully have a greater impact than the existing services available to children. Given this fact, you might want to consider a comparison of the performance of children who are receiving your intervention with the performance of children who are receiving an alternate form of intervention.

The lesson which I am learning is that, before too long, professionals in the field will not even be interested in whether intervention is better than no intervention. They will want to know whether Treatment A is better than Treatment B.

Lesson #3
Consider implementation variables as independent or control variables in your evaluation efforts.

The third lesson which I am learning pertains to the implementation of a First Chance project's "model." It is a clear fact of intervention programs that they are difficult to implement at the maximum level at all times. There will be times when a particular classroom is just not working right or when parent cooperation is exceedingly hard to obtain. Thus, natural variation exists in our intervention efforts.

It seems reasonable to assume that there should be some relationship between the fidelity of implementation of a model and the outcome performance of children, parents, staff, or anyone else. If you could document that as the

model implementation increased the outcome performance also increased, you would have a strong argument for the efficacy of your model.

A FOCUS ON DEPENDENT VARIABLES

Lesson #4
Evaluation efforts must reflect qualitative changes in children's performance.

Lesson #5
Establish the ecological validity of your intervention efforts and focus the search for outcome variables on the ecological environment surrounding the child.

The field of early childhood special education is grounded in the field of developmental or child psychology. There is emerging in the field of psychology a reassessment of what we know about how handicapped and nonhandicapped children develop. This reassessment comes from people such as McCall, Eichorn, and Hogarty (1978), Wohlwill (1973) and Scarr-Salapatek (1977). Our earlier complacency with the linear and additive nature of growth is being somewhat shaken by these theorists.

Wohlwill (1973) argues that when your research and evaluation efforts are termed "developmental," they must focus on behavioral change and the parameters of change, such as direction and rate. Scarr-Salapatek (1977) points out that such change may in fact be quite difficult to accomplish due to the very old evolutionary basis of early behaviors. The implication of such a reassessment for us in early childhood special education is that our traditional developmental instruments, such as the Bayley, the Gesell, the McCarthy, the Alpern and Boll, and others, may not be responsive to the types of changes actually occurring in young children.

There are two directions in which this reassessment can lead us. Each direction yields lessons which I am still in the process of learning. The first direction is a search for instruments which are more responsive to the qualitative, age-related changes in the children with whom we work. It is clear to me that current instruments were

not designed to measure change but instead to identify high frequency behaviors at one point in time. We must move away from this. I will address this more in my discussion tomorrow.

The second direction in which I am led is to reconsider again the importance of the parent in the family and child relationship. It seems clear that intervention which creates a positive developmental climate for these children on a long-term basis will eventually be successful, even in the absence of demonstrated early developmental changes. The outcome variable of parent and child interaction is a theoretically sound one and deserves emphasis by itself as well as with accompanying developmental changes.

FOCUS ON SIGNIFICANCE

Lesson #6
Determine the statistical and educational significance of your evaluation findings.

There is evidence in the field of educational research of a shifting focus regarding the type of significance which is judged to be necessary and sufficient. The traditional research approach argues that significance is determined when results from a research effort are judged to be statistically reliable. In other words, tests are performed on data to determine if those data could have occurred by chance at a certain level of probability.

It seems clear, judging from the recent evaluations of Follow Through (Stebbins, St. Pierre, Proper, Anderson, & Cerva, 1977) and the stated priorities of the Joint Dissemination Review Panel (Tallmadge, 1977), that significance must be statistically and educationally determined. It is not enough at this point in time to simply argue that statistically significant differences exist. A project evaluator must now argue that the differences obtained are both statistically reliable and educationally meaningful.

This shifting focus will not likely penalize the evaluation efforts of early childhood special educators. In fact, because of the small sample sizes that educators

work with, the problem which usually results is to determine the statistical significance (which is most easily influenced by sample size). I make the point that you must demonstrate statistical significance and go one step further in determining the educational significance of your results.

> Lesson #7
> An evaluation effort should be designed, implemented, and written up by a broad base of personnel in the early childhood special education program.

A final lesson that I am learning also has to do with the significance of your evaluation findings. It specifically relates, however, to the personnel involvement of your evaluation efforts. Somewhere along the way, we in the field of early childhood special education have lost sight of the fact that evaluation efforts should answer questions which we as decision makers have. I have seen too many evaluation efforts make what Mitroff and Bonoma (1978) label as an error of the third kind, or E-III. E-III is defined as the probability of conducting the wrong experiment when one should have conducted the right experiment.

I would add to Mitroff and Bonoma's suggestion the possibility that there may be an error of the fourth type, E-IV, in the field of early childhood special education. This is the probability of emphasizing in the analysis and write-up of an evaluation effort the wrong answers to the right question.

Both E-III and E-IV can be avoided or minimized by the inclusion of a broad base of personnel in an evaluation effort. They can also be avoided if early childhood special educators begin to view themselves as decision makers in the evaluation process.

SUMMARY

The field of early childhood special education is an extremely interesting and exciting profession. As time goes on and more people become involved, we are learning more about what we do and do not know. The area of

evaluation within this field is one in which early childhood special educators must actively contribute as decision makers. Throughout this conference and when you return to your projects, please keep this role in mind. My advice is to search for the decision rules that exist in the field. In this way we may find some of the long-term answers that we so desperately need.

1. Noncategorical education should not mean noncategorical evaluation.
2. Children receiving your intervention may be better off than children receiving an alternate form of intervention. An evaluation effort should make this comparison.
3. Consider implementation variables as independent or control variables in your evaluation efforts.
4. Evaluation efforts must reflect qualitative changes in children's performance.
5. Establish the ecological validity of your intervention efforts and focus the search for outcome variables on the ecological environment surrounding the child.
6. Determine the statistical and educational significance of your evaluation findings.
7. An evaluation effort should be designed, implemented, and written up by a broad base of personnel in the early childhood special education program.

REFERENCES

Bronfenbrenner, U. Is early intervention effective? A report on longitudinal evaluations of preschool programs (Vol. 2). Washington, DC: U.S. Dept. of Health, Education and Welfare, 1974. DHEW Publication No. (OHD) 74-25.

Kerlinger, F. N. The influence of research on education practice. Educational Researcher, 1977, 6, 5-12.

Kopp, C. B. Measurement issues: The mild to moderately handicapped young child. Paper presented at TADS Topical Conference on Evaluation, Chapel Hill, North Carolina, September 1978.

Kopp, C. B., & Yoshioka-Maxwell, B. The structure of interventions. Working paper, Project REACH, University of California at Los Angeles, 1978.

Lazar, I., Hubbell, V., Murray, H., Rosche, M., & Royce, J. The persistence of preschool effects. Final Report No. 18-76-07843 to the Administration for Children, Youth, and Families, Office of Human Development. Washington, DC: U.S. Government Printing Office, September 1977.

McCall, R. B., Eichorn, D. H., & Hogarty, P. S. Transitions in early mental development. Monographs of the Society for Research in Child Development, 1977, 42(3), 1-93, 106-108.

McDaniels, G. Successful programs for young handicapped children. Educational Horizons, Fall 1977.

Mitroff, I., & Bonoma, T. V. Psychological assumptions. Experimentation and real world problems. Evaluation Quarterly, 1978, 2(2).

Scarr-Salapatek, S. An evolutionary perspective on infant intelligence: Species patterns and individual variations. In M. Lewis (Ed.), Origins of intelligence. New York: Plenum Press, 1977.

Stebbins, L. B., St. Pierre, R. G., Proper, E. C., Anderson, R. B., & Cerva, T. R. Education as experimentation: A planned-variation model. An evaluation of Follow Through (Vol. 4-A). Cambridge, MA: Abt Associates, 1977.

Stock, J. R., Wnek, L. L., Newborg, J. A., Schenck, E. A., Gabel, J. R., Spurgeon, M. S., & Ray, H. W. Evaluation of Handicapped Children's Early Education Program. Washington, DC: U.S. Office of Education, Bureau of Education for the Handicapped, 1976. (Contract No. OEC-0-74-0402)

Tallmadge, G. K. The Joint Dissemination Review Panel ideabook. Washington DC: U.S. Government Printing Office, 1977. (Contract No. NIE-IA-7706)

Wohlwill, J. The study of behavioral development. New York: Academic Press, 1973.

Practical Program Evaluation: Many Problems and a Few Solutions

Owen R. White

The number of evaluation and research designs which might be brought to bear on the issue of educational program effectiveness and generalizability can virtually stagger the mind. Practically speaking, however, it is often impossible or highly impractical to utilize those designs which have been traditionally held in the highest esteem for their power and precision. Many would choose to avoid those problems by recommending that programs which cannot mount highly sophisticated and expensive evaluation operations do not attempt self-evaluation at all but, rather, employ practices already developed and thoroughly evaluated elsewhere (e.g., Cooley & Lohnes, 1976). In the face of increasing legislative and court mandates (e.g., P.L. 93-380 and P.L. 94-142), that is no longer a tenable position. Virtually all educational programs must make some attempt to demonstrate their effectiveness and, at least with handicapped children, that evaluation must reach down to the level of the individual child. The time for chastising programs for their inability to utilize traditional evaluation

---

This paper was adapted from "Curriculum-Referenced, Time-Series Program Evaluation," a presentation made at the annual conference of the American Educational Research Association (AERA) in Toronto, Canada, 27 March 78. Development of the concepts presented in the paper was supported, in part, by the Evaluation Studies Project (O. White, principal investigator), funded by the Bureau of Education for the Handicapped (#91-6001537).

strategies is past. It is now time to face the realities of the situations in which evaluations must take place and to identify reasonable evaluation strategies which have at least some potential for yielding meaningful results within the practical constraints of those situations. It shall be the purpose of this chapter to at least initiate that process.

PRACTICALLY SPEAKING

From the standpoint of evaluation resources, educational programs might be divided into two categories: (1) special, heavily funded development or demonstration programs which can afford special evaluation services and facilities; and (2) ongoing service programs which must conduct their evaluations with existing personnel and facilities. For purposes of these discussions, I shall concentrate on the needs of the latter. Moreover, I shall address myself primarily to the needs of programs serving the severely and/or profoundly handicapped, since they face some very special evaluation problems of interest.

When designing an evaluation strategy for such programs, at least the following factors must be considered:

1. PERSONNEL AND OTHER RESOURCES. Most evaluations must be conducted with regular personnel. Those people are not likely to be highly trained in evaluation or research strategies, nor are they likely to have a great deal of extra time to devote to program evaluation--especially if the results of those evaluations are not immediately useful to them in their primary roles as educators.

2. ETHICAL PUPIL SELECTION. Few programs have any real choice over the selection of the pupils they serve. At best, a few criteria for acceptance into the program will be established so all children in a given program will share some basic characteristics, but it will not usually be possible to prevent children from entering the program simply to meet the requirements of some evaluation strategy.

3. THE NEED TO IMPROVE PROGRAMS. Most service programs will not be interested in developing general programs solely for the sake of widespread dissemination.

Rather, emphasis is placed on the continuous development and refinement of the best service package possible for the benefit of pupils currently in the program. If something goes wrong, even for one child, teachers and administrators will want to change the program immediately. Strict timelines defining preplanned periods of evaluation during which conditions are held "constant" may not, therefore, be realistic.

4. INDIVIDUAL EVALUATIONS. Traditionally, overall program effectiveness has been evaluated by assessing all children in the program with the same basic instrument. Since the performance of all children is held relative to the same criteria, scores can be grouped and aggregated for easy analysis. Laws now exist, however, which require the evaluation of pupil progress in terms of individually established goals and objectives. That means that evaluation strategies must be devised which can somehow combine those "individualized" assessments in a meaningful manner, or to allow staff extra time to complete two sets of assessments--one for the assessment of individual progress and the other for the assessment of overall group progress.

5. THE SEVERELY/PROFOUNDLY HANDICAPPED. Several special problems are encountered when attempting to evaluate programs serving the severely and profoundly handicapped. First, they represent a very low-density subgroup of the general population (that is, there really aren't very many of them). A school district with 20,000 pupils, for example, might expect to serve only 150 to 200 severely or profoundly handicapped children between the ages of three and 21--the age range mandated by current laws. With so few children, there is very likely only one program in any given area which serves those children, thereby limiting the availability of the "control" or "contrast" groups necessary for many types of evaluation strategies. The sparsity of programs also exacerbates the problem of pupil selection--virtually all severely or profoundly handicapped pupils must be served by the same program, regardless of their specific disabilities or personal characteristics. That raises the second major evaluation problem: the subgroup of severely and profoundly handicapped people is extremely heterogeneous (that is, they aren't very much like one another). There are eight major handicapping conditions

recognized by the Bureau of Education for the Handicapped. Most severely or profoundly handicapped children are handicapped in more than one way, however, raising the number of possible categories to at least 46--without even considering the actual level of deficit or the person's age. In our hypothetical district of 20,000, then, it is quite unlikely that more than two or three severely or profoundly handicapped pupils could be found who matched each other reasonably well on even the most obvious educationally relevant variables. As will be demonstrated below, those low-density/high-heterogeneity characteristics of the severely and profoundly handicapped population virtually nullify the possibility of utilizing the vast majority of traditional strategies for the evaluation of programs serving such pupils.

THE BASIC QUESTIONS

Regardless of the specific approach employed, the overall purpose of any program evaluation is to answer two basic questions:

- Did the pupils served by the program in question achieve "meaningful" levels of performance and/or demonstrate "acceptable" rates of progress?

To answer this question one must consider the general goals of the program, whether they are appropriate for the pupils involved and how progress toward (or mastery of) those goals will be measured. These are serious issues which cannot be taken lightly. If we do not know what we want a program to accomplish or how we are going to measure that accomplishment, then certainly no reasonable program evaluation is possible. Nevertheless, these issues are not unique to the evaluation of service programs nor to the application of any particular evaluation strategy. Aside from a few comments presented later, therefore, we will concentrate on the second basic evaluation question.*

---

*For a more complete discussion of setting individualized objectives, see White (1980).

- Is there any reasonable evidence to indicate that the program in question was at least partially responsible for the observed performance or progress of the pupils?

It is not sufficient to simply say that the pupils in a program progressed satisfactorily. We must also demonstrate that their progress was due, at least in part, to the program in question. In all cases, that will amount to forming some reasonable estimate of how well the pupils would have performed if they were not included in the program, comparing that estimate with measures of how well they did perform, and then demonstrating that any differences between those figures cannot be reasonably attributed to simple "chance" or "accident." The major differences among various sorts of evaluation strategies lie in how they go about forming those estimates and making those comparisons.

GROUP DESIGNS

The most favored group design evaluation strategies use "control" and "treatment" groups. The simplest design is to select two comparable groups of pupils. One group (the treatment group) is enrolled in the target program. The other group (the control group) is enrolled in some other program or not served at all. If the treatment group demonstrates consistently higher overall rates of progress, then we conclude that the program was helpful. If the two groups end up pretty much the same, we conclude that the program was of no real benefit. In situations where it is difficult or impossible to find two or more equivalent groups of pupils, groups which are systematically different from one another might be used for what is called regression-projection or regression-discontinuity analysis. Finally, if the whole notion of two or more reasonably comparable groups is impossible, then it has been suggested that meaningful group evaluations might still be conducted by using the standards associated with norm-referenced tests in place of the contrast or control group (Tallmadge &

Horst, 1976). Many programs, however, will encounter difficulties with all of those approaches.*

EQUIVALENT CONTROL GROUPS

When dealing with programs serving normal or mildly handicapped children, many districts will find themselves with a sufficient number of pupils to consider the use of equivalent control group evaluation designs--assuming, of course, that there are no ethical problems in assigning different children to different types of programs. If the target population is small, however (as is the case with the severely or profoundly handicapped), then the program in question may be the "only game in town." Direct comparisons with other service options in the same general area will be impossible and the option of simply leaving some children out of programs is usually untenable. Locating programs in other service areas which might be used as "controls" raises the issue of "natural selection" (i.e., if there are differences in program effectiveness, perhaps they are simply due to the same factors which led different people to live in different areas) and, from a practical standpoint, political issues arise when trying to design an evaluation study for the purpose of showing that one district's program is better than another's. So how do we find two or more groups of children that are essentially similar?

The obvious solution is to assign children being served by the same basic program to different treatment plans--provide some services to some children and other services to other children. Assuming that all of the treatment options are "reasonable" (i.e., they represent the types of programs which the facility might provide anyway), such a design would allow a school or agency to find out what really works and what doesn't. That will

---

*There are many other variations of group-statistical evaluation models. An excellent overview of variations used with educational programs may be found in Tallmadge and Horst (1976). A more complete discussion can be found in Winer (1962).

only be a viable option, however, if the agency is serving a relatively large number of children who are very much like one another or who can at least be reasonably matched in pairs. For the reasons discussed earlier, that is not likely to be the case with many programs.

REGRESSION-PROJECTION AND REGRESSION-DISCONTINUITY MODELS

If pupils are selected for inclusion in a program on the basis of some absolute cutoff score on a test (e.g., all children scoring below a certain level on an examination of their motor skills might be offered physical therapy), then it may not be necessary to find directly equivalent control groups. The relationship between pretest and posttest scores for children not included in the program might be used to estimate the expected relationship between the same tests for those who are included in the program. The actual posttest scores of the treatment group are then compared to their predicted scores to see if the treatment had any effect. The two procedures are illustrated, briefly, in Figures 1 and 2. If the overall effect of the program should be about the same for all children, regardless of entry score, then the regression-discontinuity model is appropriate (Campbell & Stanley, 1963). If the treatment is likely to be more or less effective with a child in accord with his initial level of deficit, then the regression-projection technique is more appropriate (Tallmadge & Horst, 1976). In either case, the assumption that the performance gains of higher-level children can be used to estimate the expected gains of lower-level children may be unwarranted.

In order to use the gains of high-level children to estimate the gains of low-level children, we have to assume that the factors which affect those gains are "smoothly distributed" across both groups of children. For some variables, that may be true. Pretend, for example, that there was a direct relationship between the rate with which children learn mathematics and their intelligence quotients. Perhaps a student with an IQ 20 points higher than another student will be able to learn mathematics in one-half the time. If that were true, and since all children have <u>some</u> IQ, we could use the difference in IQ between <u>our</u> two groups to estimate how

FIGURE 1
Regression-Discontinuity Model

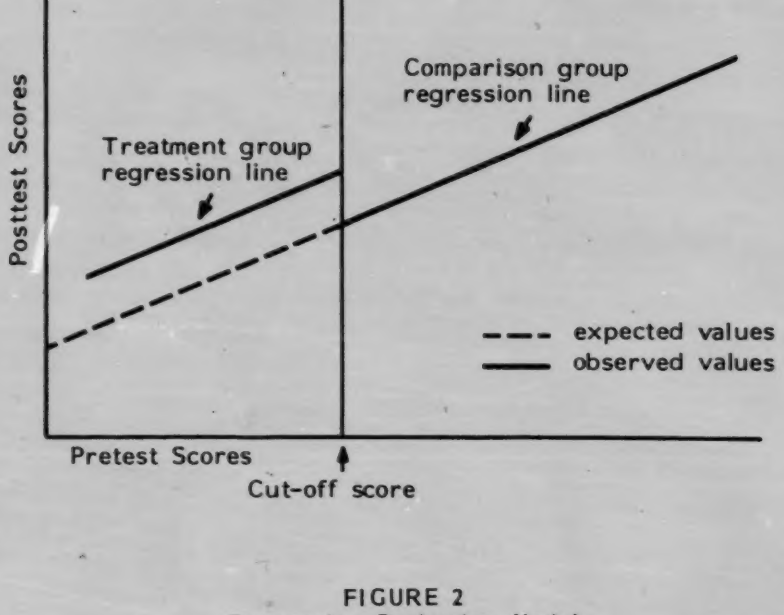
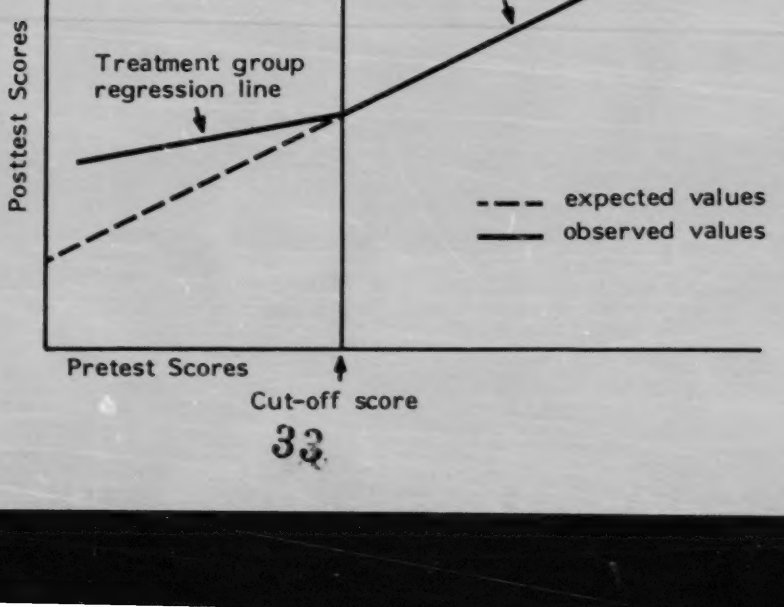


FIGURE 2
Regression-Projection Model



rapidly the treatment group should learn. If they actually learn faster than that estimate, we could conclude that our program was effective. Many variables are not similarly distributed across groups, however. In measuring the impact of a program for teaching children without arms to use artificial limbs, we certainly cannot use the fine-motor development of children with arms to estimate expected rates of progress. It would be foolish to assume that the relationship between initial performance and performance gains in non-crippled children was at all similar to the expected development of children without arms. The same might be said for the developmental characteristics of blind children, deaf children, emotionally disturbed children, or even bilingual children. Those children are qualitatively different than their nonhandicapped peers, not just quantitatively different. There are many types of educational programs where the regression-discontinuity and regression-projection models would be inappropriate.

NORM-REFERENCED MODELS

In cases where we cannot find or assess a reasonable control or contrast group of our own, it has been suggested that the standards associated with norm-referenced tests can be used as the basis of comparison (Tallmadge, 1977). In essence, we use pupils assessed by other people (i.e., the developers of the norm-referenced test) as our "control group." If the norm-referenced test is a valid measure of how pupils "usually" develop, then it can be used to estimate how quickly our pupils should develop if they were exposed to the same type of programs those other pupils experienced. If our pupils develop more rapidly (e.g., actually increase in IQ, or whatever the measure is), we can conclude that our program is more effective than the alternative programs experienced by the "typical pupils" used to norm the test. Still, as Horst, Tallmadge and Wood (1975) point out:

> The validity of the [norm-referenced] model rests on the assumption that the achievement status of a particular subgroup remains constant relative to the norm group over the pre- to posttest interval if no special treatment is provided. Empirical support for this assumption is minimal. It is conceivable that some subgroups would move up and others move down in the normal course of events. When the norm group is like the treatment group, the plausibility of the underlying assumption is greatly enhanced; thus, for example, norms for gifted children would be best for assessing a project serving such pupils (p. 72).

Unfortunately, special norms for selected subgroups of the general population are virtually nonexistent, especially for groups like the severely/profoundly handicapped. The validity of existing norms are in even greater doubt with such children than with most subgroups, since few if any severely or profoundly handicapped children were ever included in most norming samples. As a result, the authors and developers of the most widely used normed assessment instruments usually warn users not to employ their tests with handicapped children and don't usually provide values in their raw-score conversion tables for children functioning below a certain level.* Unless the children in your program are

---

*Some attempts have been made to develop norm-referenced instruments for special populations like the severely/profoundly handicapped. Since it is generally difficult to find very many such children located in any given area, however, those attempts have usually failed. Sailor and Mix (1975) could find only 300 severely and profoundly handicapped people for their norming of the TARC system; and those people ranged in age from three to 21 with all types of handicapping conditions! Obviously, that is not a large enough sample to develop meaningful norms for all of the population subgroups. Until a way is found to collate information on larger groups, the development of norm-referenced instruments for many subgroups will remain only a dim hope for the future.

very close to being "normal," therefore, the use of norm-referenced test data for program evaluation is questionable.*

AND THERE'S STILL THE INDIVIDUAL...

Beyond the problems associated with control groups, regression-discontinuity/projection models or norm-referenced evaluations, there is still the problem of accounting for individual differences. No overall group summary of pupil progress will meet the legislative mandate of P.L. 94-142 or similar statutes. The value of group estimates and summaries should not be underestimated--they will always be useful for at least some general administrative decisions--but if we must evaluate individual pupils anyway, it would certainly be to our advantage if we could find a way of using those data for overall program evaluations as well.

"INTENSIVE" OR "SINGLE-SUBJECT" TIME-SERIES EVALUATION DESIGNS

Rather than conduct an "extensive" examination of many pupils, several people have suggested the "intensive" evaluation of the performance patterns of individual pupils over time (i.e., "single-subjects" in a "time-series"). Each pupil could be exposed to a variety of programs and, by comparing his performance patterns in one program with his own performance patterns in other programs, we might discover which programs work best. Each pupil acts as his own "control," so we don't have to worry about finding similar pupils to act as separate controls in another group. Finally, by running several such studies on most or all of our pupils, we can determine whether the program works equally well with all pupils, only with certain types of pupils, or not at all. We meet both our individual-pupil evaluation requirements

---

*The use of most standardized criterion-referenced instruments is equally questionable for many of the same reasons (White & Haring, 1978; Tallmadge & Horst, 1976).

and our overall program evaluation requirements! It sounds good and has certainly been a favorite approach of many evaluators for several years (e.g., Sidman, 1960; Herzen & Barlow, 1976). Still, it's no panacea--nothing could be that simple!

There are many approaches to single subject evaluation, the most common of which are the "withdrawal" or "ABA" design, the "reversal" design, and the "multiple-baseline" design.*

ABA DESIGNS

In the ABA design, the pupil's performances are first assessed for several days or trials under a "no program" condition (the first "A" phase) to determine how well he or she progresses without the program.** The program is then introduced (the "B" phase) and several additional assessments are conducted to determine whether the pupil's performances or rates of progress improve. We might stop our evaluation at that point, but one potential problem remains. The pupil's performance might have improved dramatically when the program was introduced, but it is impossible that some other, non-program-related factor was really responsible for the improvement. The pupil is, after all, older. Perhaps the improvement was a result of simple maturation. Time has also passed--we're further away from the beginning of school, or closer to Christmas, and perhaps those changes were really responsible for the improvement. In order to demonstrate

---

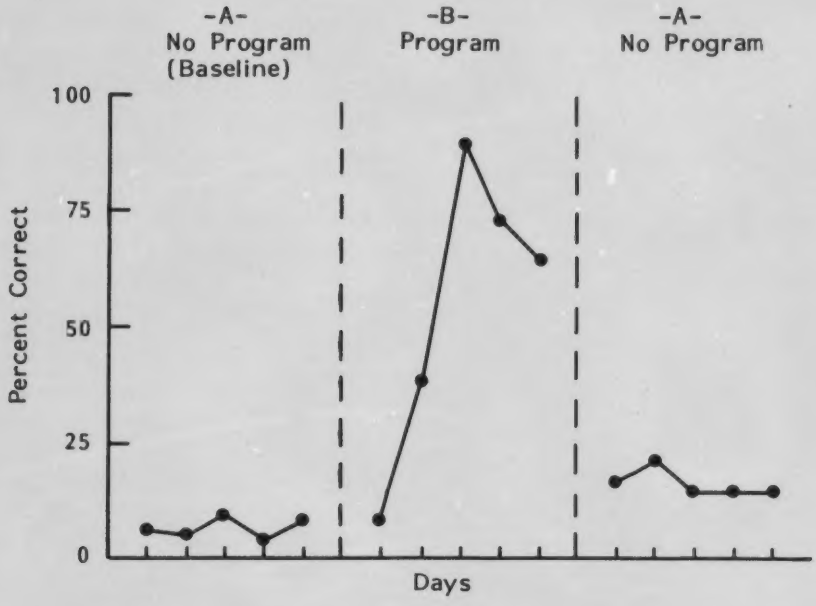
*For a more complete discussion of single-subject design options, consult Sidman (1960), Hersen and Barlow (1976), and Kratochwill (1978).

**Actually, it is usually recommended that the "no program" or "A" phase represent some alternative condition rather than no program whatsoever. It is usually more convincing to compare the effects of a new program against a reasonable alternative than to compare the new program against the complete absence of a program.

that the improvement was not due to those other factors, therefore, we should terminate the program and see if the pupil maintains his or her improved performances without it (the second "A" phase). If, indeed, the pupil progresses much better in the program ("B") phase than in either of the non-program ("A") phases, it might be concluded that the program itself was effective--at least with that child (see Figure 3).

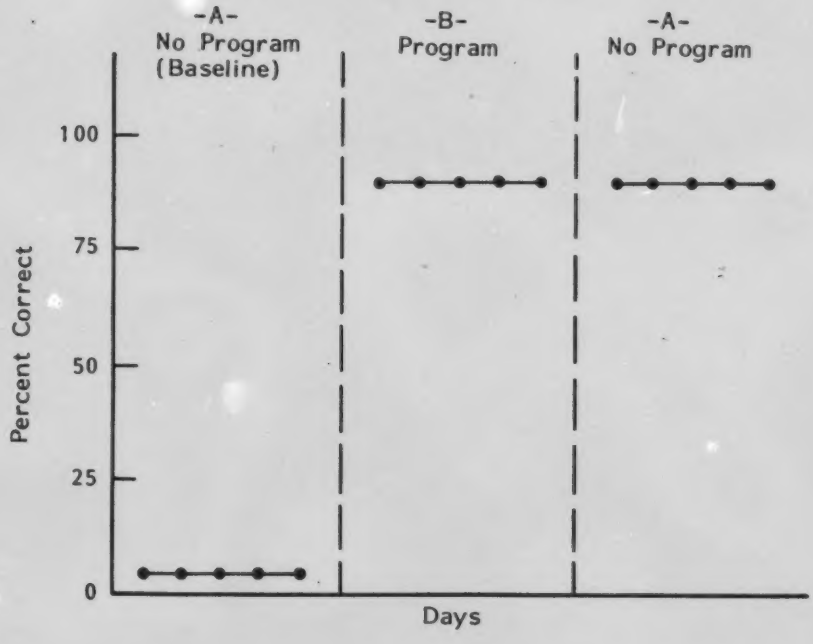
The ABA design can, of course, be extended to include many applications and withdrawals of the program and/or to test a series of different program options (e.g., ABABCBCAC...). Each up or down shift in performance when the program is applied or withdrawn simply adds to the confidence we can place in our estimate of the program's effect. If everything goes well, the ABA design can prove very convincing. Problems with the design can arise in actual practice, however.

FIGURE 3
ABA Design Showing Performance
Returning to Baseline Level



Not all learning is "reversible." That is, the behavior might improve or develop during the first "B" (treatment) phase, but fail to return to the baseline or pre-treatment levels when the program is terminated (see Figure 4). After a baseline to establish that the child does not know certain skills, a program is implemented to help her acquire a desired skill. Later, the program (treatment) is removed and, not surprisingly, the child continues to demonstrate the skill. Educationally, that's desirable. It would be terrible if all children "forgot" their skills or "unlearned" new behaviors as soon as each program was terminated. Experimentally, however, the failure for the behavior to return to some lower level detracts from the power of the study to demonstrate clearly that the program itself was responsible for the observed change in performance. As mentioned earlier,

FIGURE 4
ABA Design Showing Performance
Maintaining Program Level



some other event, unknown to the experimenter, might really be responsible for the change and only coincidentally occurred at the same time the program was introduced. Perhaps, for example, this same child just happened to watch Sesame Street and learned just the right skill at the very time when the treatment program was supposed to be introduced. That may seem far-fetched but it's possible, just as its possible that the child could have imitated a playmate performing the same skill. If the performance of the child had returned to baseline (preprogram) levels when the program was terminated, then our confidence in the program would be justified; but since the behavior did not return to baseline, we're left with a lingering doubt. If the ABA design is used, therefore, it is wise to select behavioral targets which are likely to improve only as long as the program is in effect. If there is reason to believe that any change produced by the program will be permanent, one of the other designs outlined below will be more appropriate.

REVERSAL DESIGNS

Reversal designs are very much like ABA designs with one major exception. Instead of simply withdrawing or terminating the treatment program after the behavior has been improved, the program is changed in a way which actively promotes the reversal or deterioration of the performance. If a study were designed to show how a child's in-seat behavior can be improved by awarding points for staying in his seat, the power of the points could be further demonstrated by awarding points for out-of-seat behavior once in-seat behavior has been established. If the points increasing the child's in-seat behavior, they should prove equally powerful in reducing his in-seat behavior. Luckily, very few actual examples of reversal designs can be found in the educational literature. Most people take a dim view of actually teaching a child inappropriate behaviors simply to demonstrate the power of an educational procedure or program. If the behavior in question is of little practical significance to the child, or if one can be quite certain that no real damage might be done by confusing the child with conflicting programs, then the reversal design can be very powerful for demonstrating program effect.

Great care must be exercised, however, in addressing the ethical implications of such procedures.

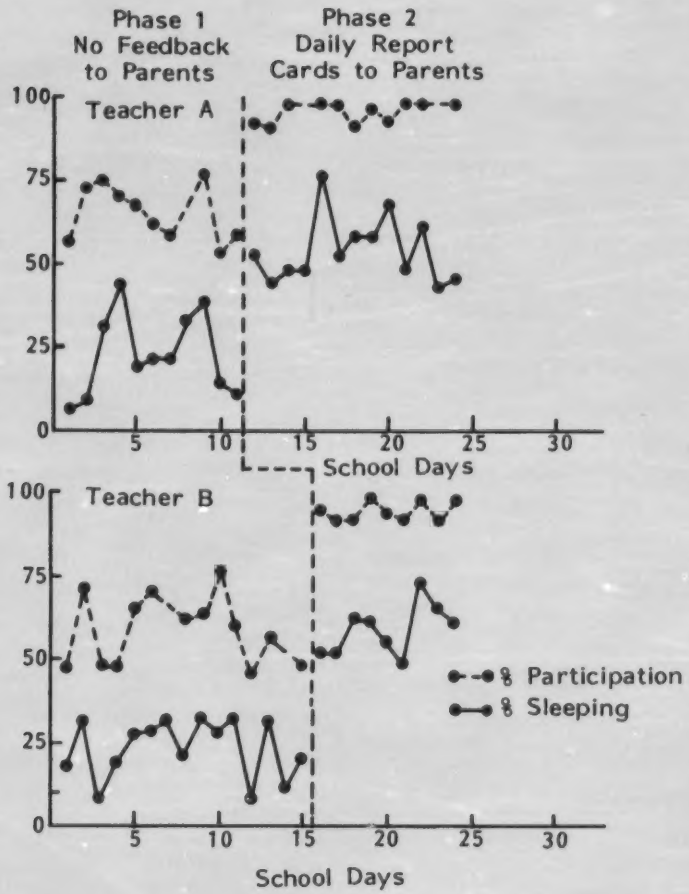
MULTIPLE BASELINE DESIGNS

If a behavior is irreversible (i.e., won't get worse again when the program is terminated) or if it is undesirable to reverse the behavior (either by terminating the program, as in the ABA design, or by actively working to reverse the behavior, as in the reversal design), the possibility of "coincidental" events which might actually have been responsible for the behavior change can be tested by trying the program with many different behaviors or children at different points in time.

An example is provided in Figure 5. Lahey, Gendrich, Gendrich, Schnelle, Gant, and McNees (1977) wanted to see whether a system of daily reports to parents could improve the resting and sleeping habits of a group of kindergarten children during the appropriate period of the school day. If the procedure appeared to work, they didn't want to terminate it (to achieve a full ABA design) and they certainly didn't want to actively encourage poor resting behaviors (as they would with a reversal design). In order to help rule out the possibility that any changes they observed might be due to some coincidental event (e.g., the approach of a vacation), therefore, they decided to implement the procedures at different times in different classrooms. As can be seen, the behaviors improved dramatically in each class only when the procedures were introduced. Since the probability that "coincidental" events just happened to occur at exactly the right moment in both classes is low, the evaluators can be reasonably confident that the change was due to the reporting system they were testing.

Multiple baseline studies can be run using the same behavior and different children or groups of children, or they can be run using only one child and several different behaviors. Some care must be taken in the selection of subjects and behaviors, however, to make sure that a change in one behavior or subject does not influence changes in the other behaviors or subjects. For example, if studying the usefulness of praise for increasing a child's interactions with other children, it is likely that the behavior of many children in the class will

FIGURE 5
A Simple Multiple-Baseline Study
Using Two Classrooms



Source: Adapted from Lahey, Gendrich, Gendrich, Schnelle, Gant and McNees (1977).

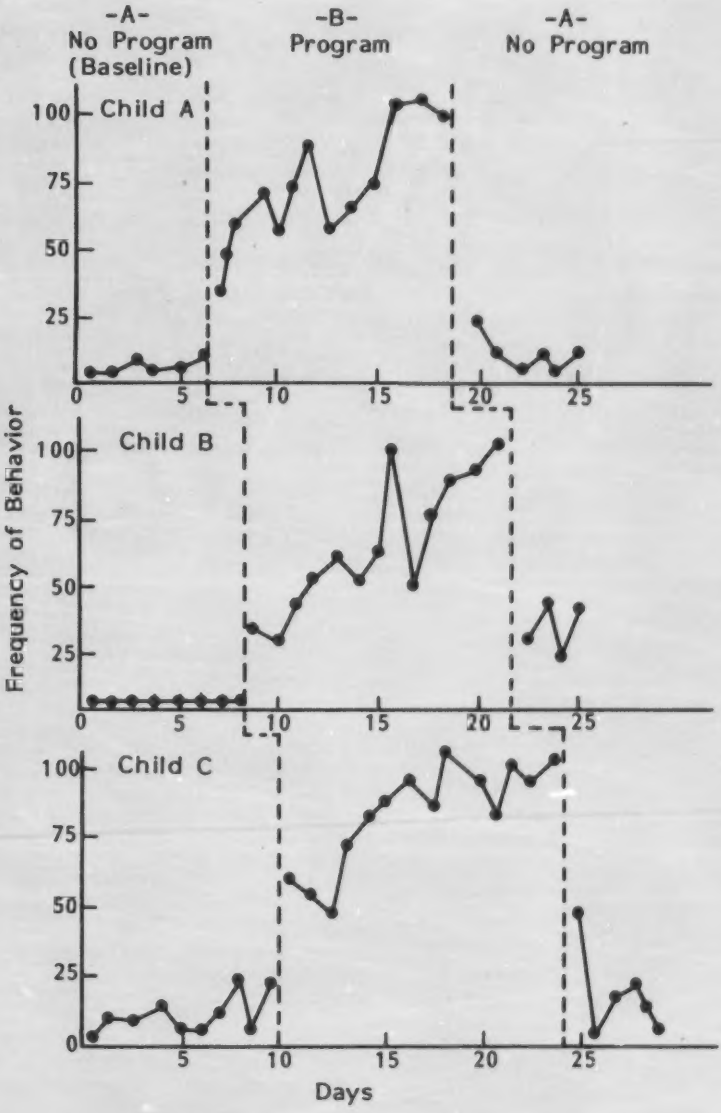
change even if praise is only given to one child. After all, the subject of the study would have to interact with someone. Rather than simply selecting new children at random for inclusion in such a study, therefore, it would be necessary to look at each child's performance records and select only those children whose behavior did not seem to change much when the program was implemented earlier with other children. Similarly, when studying several behaviors in one child, care must be taken to select behaviors that are not likely to affect one another.

If some care is exercised in the selection of assessment targets (i.e., children, groups and/or behaviors), the multiple baseline approach to evaluation can provide powerful evidence of program effect while avoiding some of the programmatic and ethical issues surrounding the ABA and reversal designs. Indeed, the idea of staggering the times when programs are implemented with different behaviors and/or children is so effective in dispelling any doubts concerning coincidental events that many researchers suggest using the multiple baseline approach even if ABA designs or reversal designs can be used. Even more powerful designs can be achieved by blending and combining approaches, as illustrated in Figure 6. Still, multiple baseline approaches to evaluation don't solve all the problems one is likely to encounter. Establishing the baseline itself can sometimes be very difficult.

ESTABLISHING THE BASELINE

The validity of multiple baseline studies (or, for that matter, virtually all other types of time-series studies) rests on the manner in which baseline or pretreatment performances are assessed. Unless some confidence can be gained that preprogram performances have been meaningfully and reliably determined, the rest of the evaluation will be useless. Unfortunately, establishing an adequate baseline will prove difficult in many applied settings.

FIGURE 6
A Combination of ABA and Multiple Baseline Evaluation Designs



WITHHOLDING TREATMENT

The most common approach to establishing a preprogram baseline is to withhold the treatment or program until preprogram assessments have been completed. To some extent, that presents the same ethical concerns encountered when trying to justify a "no-treatment" control group. If baselines can be established in a few days or weeks, however, perhaps the problem will not be of great concern. There may even be reasons why the program cannot be implemented right away: teachers might find it difficult to begin all programs at once, and, at least in some cases, a child might not be ready for a particular program. If so, then the "natural" delay in program implementation might be used as an excuse for taking the time to establish a preprogram baseline. Each of those rationales creates problems of its own, however.

If a teacher is unable to begin all programs at once, it is also quite likely that he or she will not have the time to collect reliable baseline information. Even if someone else can be found to collect the data, comparing the effects of a program with the effects of a "nonprogram" may not be very interesting. If a child does better when we do something than when we do nothing, so what? That will provide no information about the power of the program as compared to reasonable alternatives.

Similarly, if a program is withheld or delayed because a child is not yet "ready," then data collected during that period could not really be meaningfully compared to data collected during a period when the pupil is ready (i.e., presumably a period during which the program would be initiated). Obviously, any differences in the pupil's performances between the two periods might be due only to differences in "readiness" and have nothing to do with the effects of the program per se.

It seems apparent, therefore, that "natural excuses" for withholding treatment during baseline periods can create more problems than they solve. Either the educator must face the ethical concerns of withholding treatment head-on, or seek other ways of establishing an estimate of the pupil's preprogram performances.

HISTORICAL DATA

In some cases, preprogram baseline performances can be established by analyzing performance data collected prior to the time when the program was available (e.g., prior to the time when a pupil is referred to the program). The availability of such data is generally rather low, however, and even when some data are available, their reliability might be questionable.

As an alternative, many evaluators have attempted to estimate preprogram rates of progress on the basis of the pupil's age, level of performance at the time when the program was begun, and other demographic information. Assume, for example, that a program is about to be implemented with a pupil who should have already developed the skill in question, but who has only achieved 30 percent of the criterion established for skill mastery. Assume also that the pupil should have begun to develop the skill one year earlier. We might, therefore, estimate that the pupil's preprogram rate of development was 2.5 percent per month (30 percent divided by 12 months =). Now, if the pupil's actual rate of development during the program is greater than 2.5 percent per month, we might assume that the program was effective. Obviously, we would be going out on a limb with all those assumptions.

First of all, we don't really know when the pupil actually began to develop the skill. Assuming that development began when most children begin to develop is particularly tenuous in light of the fact that the child is currently delayed. If he's delayed in completing skill development, would it not be reasonable to assume that he was also delayed in initiating skill development? Perhaps skill development actually began only one month ago, rather than 12 months ago. If so, his rate of progress during the program might actually be less than his rate of progress before the program (see Figure 7).

Second, many skills are acquired in an "exponential" fashion--each new step in skill mastery is easier than preceding steps. If that is the case, then the pupil's absolute rate of progress will be continuously increasing, and the rate of growth for any period will be greater than the average rate of growth for any preceding period. What looks like a program effect might, therefore, be just the natural progression of events (see

FIGURE 7
Potential Overestimate of Child's Progress
When Initiation of Skill is Incorrectly Assumed

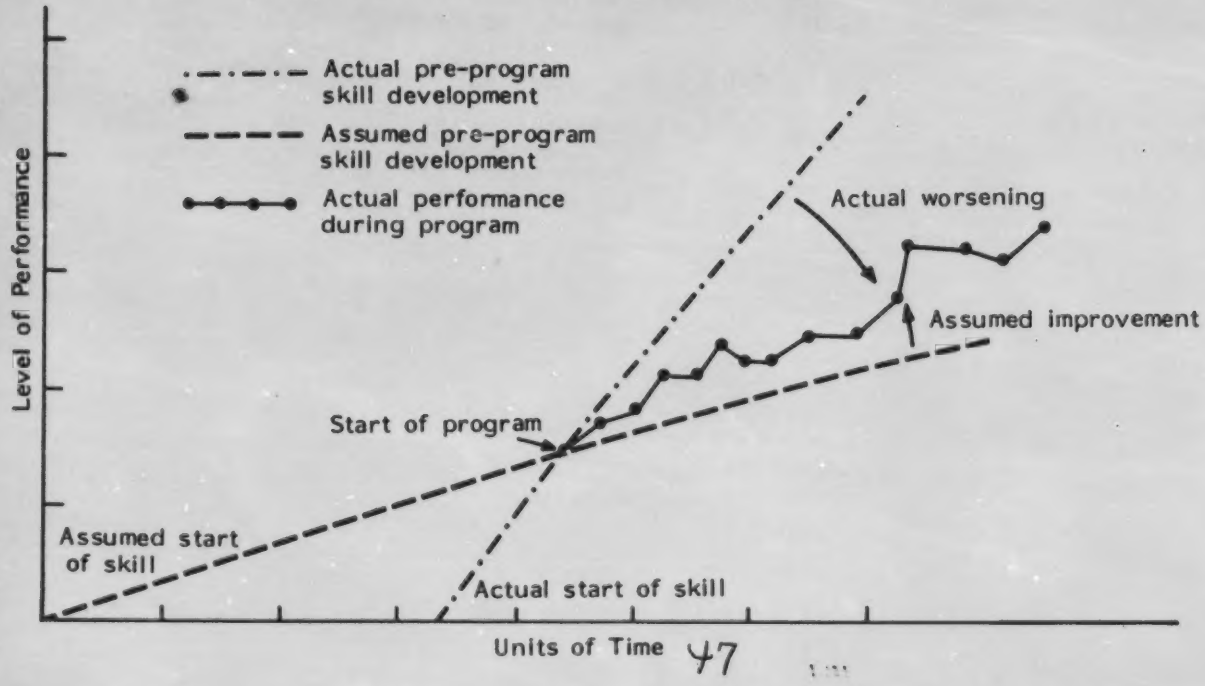


Figure 8). If several assessments of pupil progress are made during the program period, it should be possible to detect a continuously accelerating trend in rates of progress and to take that trend into account when evaluating possible program effects. Misinterpretations of such data are, however, made with some regularity.

Several methods for obtaining reasonable estimates of preprogram progress have been suggested (e.g., Bryk & Weisburg, 1977; White & Haring, 1978), but all depend in some part on group analyses of relationship between age, entry performance, and rates of skill development. For such analyses, of course, a common method for assessing progress in relatively large groups of pupils would be required--a problem I will address a little later in this chapter.

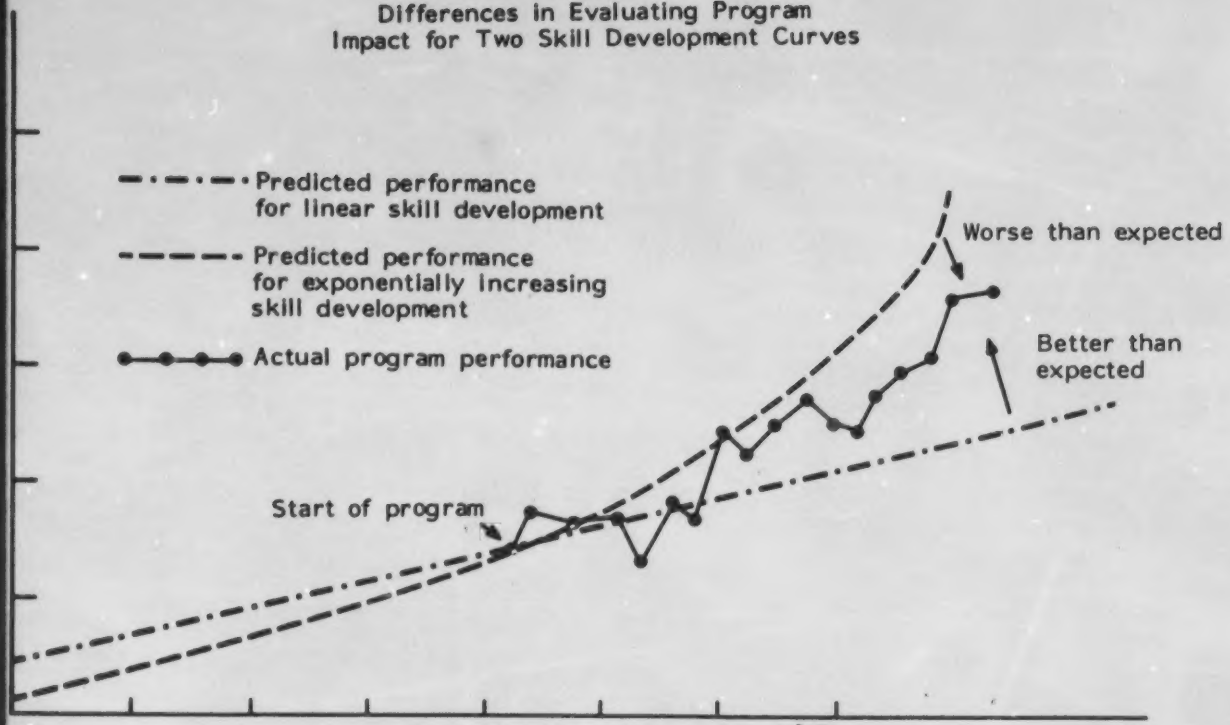
TREATMENT VS. TREATMENT

As mentioned earlier, most applied educational programs are in a constant state of change. As individual or group programs fail to produce expected results, programs are modified and refined. If regular data are being collected to reflect day-to-day pupil progress, it is sometimes possible to document the effects of those changes with some precision. Events which appear to improve one child's program might then be tried with other pupils, thereby creating a natural multiple baseline study. Of course, such an approach to program evaluation amounts to capitalizing on one's failures--demonstrating that what one thought would work did not and that some alternative program did. As the overall program improves, the opportunity to employ such a strategy decreases. If program failures are encountered (or, if new ideas arise for ways of improving already successful programs), the use of treatment vs. treatment multiple baseline studies to document program effect is a nice approach, but it is generally not possible to rely on such an approach to meet primary program evaluation needs.

WHAT ABOUT THE OVERALL PICTURE?

While not without problems, multiple baseline studies for the evaluation of individual educational programs do

FIGURE 8
Differences in Evaluating Program
Impact for Two Skill Development Curves


Predicted performance for linear skill development
Predicted performance for exponentially increasing skill development
Actual program performance
Worse than expected
Better than expected
Start of program
Units of Time

offer some promise. When properly conducted, they can provide very convincing evidence of program impact and (when several behaviors and/or pupils are studied) they can provide information concerning the generalizability of program effectiveness across different program targets. Perhaps just as important, such studies provide information on a day-to-day basis which is far more useful to most educators in continuously refining their programs than the type of "summary data" usually provided by large group-statistical approaches to evaluation. Due to the problems mentioned above, however, it is not likely that all of every pupil's program will be amenable to such studies. There may even be some parts of the program or some pupils where no multiple baseline studies can be performed. That raises the possibility that the overall impact of the program as a whole might be misrepresented.

We might avoid, or at least minimize, the possibility of misrepresenting a program's overall impact by making sure that the targets of study are selected at random from the set of all possible pupils and instructional programs. Unfortunately, however, the set of all instructional programs conducted within a school or class is still not likely to represent all areas of potential concern. Instructional programs are generally established to meet what is considered to be each pupil's greatest needs, but that does not mean that pupils do not have other needs. A speech impaired preschool child may require special attention in the areas of expressive communication, but that does not mean that preacademic skills, fine and gross motor development and social/self-help skills are of no concern. Special programs may not be implemented to deal specifically with those needs, but if they are not, it must be assumed that they will develop "naturally," or at least not be directly impaired by the focus of programs in the area of communication. Expressing a similar concern, Tallmadge (1977) pointed out:

> It is occasionally possible to conduct a credible evaluation without an empirically based estimate of what conditions would have been without the intervention. Where a totally new topic is introduced into a curriculum [or to a child], for example, it is not necessary to demonstrate

> empirically that a control group [or the same child], not exposed to the topic, [would have] learned less of it.... On the other hand, it should be demonstrated that the new learning was in addition to, not in place of, what was learned previously [or would have been learned] (p. 18).

If, therefore, special multiple baseline studies are conducted to demonstrate the impact of a program on only some subset of the behaviors which might be of concern with the pupils involved, some other means must be employed to demonstrate that the programs did not detract from development in those other areas. To meet that evaluation need, we must renew our search for some general measure of overall pupil development and progress.

CURRICULUM-REFERENCED ASSESSMENTS

In the absence of any meaningful, fully validated national norms which describe how pupils should develop in each major area of potential concern, many educators have turned to "curriculum-referenced" assessment instruments. According to English and English (1958), an assessment instrument has curricular validity to the degree that "test items are representative of the knowledge, skills or attitudes supposed to be learned in a given course of study (p. 575)." In order to construct a valid measure of a child's progress through a program, therefore, perhaps a list could be compiled of the program's major objectives and concerns. As the pupil mastered each objective or demonstrated competency in each area of concern (whether or not direct instruction in that area was actually provided), that concern could be checked off the list. The specific impact of program elements could still be evaluated through careful, detailed, multiple baseline studies of selected instructional programs, and the overall checklist of concerns could be used to provide assurance that progress noted in those studies was not gained at the expense of progress in other areas of concern.

Individual lists of concerns could be compiled for each pupil, but the effort required to compile and validate such lists would make that task almost insurmountable. Generally, it will be necessary for a program to attempt the compilation of a uniform list of concerns which is then used to evaluate the overall progress of all (or at least most) of the pupils in the program. Is that reasonable? Many of the arguments presented earlier in this chapter were directed at the issue of "individualization." Now it seems that I am arguing for generalities. One must bear in mind, however, that all of the objectives listed will be at least generally "program-relevant" and that all of the children in any given program will share at least some needs in common. With some children, many needs on the list will have already been met (e.g., many children may already be toilet-trained before entering the program), but that does not detract from the importance of the skill or the importance of finding those children who have not yet mastered the skill. Similarly, some children in a program will never master certain of the "common" program objectives (e.g., some orthopedically handicapped children may never walk), but that does not detract from the value of knowing which children lack the skills, and certainly the need to at least <u>try</u> and develop the skill (in some form) still exists. It will frequently be difficult to identify objectives which can be adapted meaningfully to reflect the needs of individual children and some compromises will be necessary. It is possible, however, and as will be outlined below, many advantages can be realized through the attempt (see also White, Reavis, & Kunzelmann, 1978; White, 1980).

SCALING

If a child's score on a curriculum-referenced checklist is to have any meaning, several basic scale requirements must be met. First, the scale must be constructed in a manner that allows at least ordinal data to be collected--that is, a higher score on the scale must reflect a higher status or level of development. If the objectives are listed in a strict hierarchy (i.e., each objective in the sequence must be mastered before the next objective can be completed), then higher scores will always mean more advanced development. In most cases,

however, strict and complete hierarchies of skills are not possible (e.g., most children will learn how to walk before they learn the alphabet, but a crippled child might learn those skills in reverse order). If strict hierarchies cannot be assured, then objectives weighting becomes important. Each objective must represent a developmental concern which is essentially as important as each other objective. Then, if a pupil advances from a score of 10 to a score of 20 on a scale, we can be relatively confident that he has, indeed, improved and not simply learned some skills at the expense of other, possibly more important skills. Similarly, if the scale is ordinal (either through hierarchies or through equal weighting of items), meaningful comparisons between different pupils or groups of pupils will be possible without undue concern for the specific objectives each pupil has mastered. Finally, if it is reasonable to expect each skill to take about as much time to master as every other skill on the checklist, reasonably accurate projections of expected rates of progress can be made. It will then be possible to use the checklist itself to provide data for a type of "large scale" multiple baseline study of overall program impact with individual students or whole groups of students.

IT CAN BE DONE!

Despite the obvious difficulties in devising a reasonable curriculum-referenced instrument, it is possible. Several such instruments already exist for programs serving moderately, severely and profoundly handicapped children* and a new instrument is currently being developed by the Consortium on Adaptive Performance Evaluation (CAPE)** which promises to be particularly adaptable to

---

*See Walls, Werner, Bacon, and Zane (1977). Another listing of scales, specifically for use with the severely/profoundly handicapped, was compiled by Lestar Hazel and is contained in the Experimental Education Training Program, N. G. Haring (Ed.), Experimental Education Unit, CDMRC, WJ-10, University of Washington, Seattle, WA 98195.

**For information on CAPE, contact Dr. Dale Gentry, College of Education, Department of Special Education, University of Idaho, Moscow, ID 83843.

the unique needs of multiply handicapped pupils. Not all available instruments are equally well suited to each type of program, of course, but many instruments are accompanied by computer programs which enable the precise analysis of any additions, deletions or revisions desired by a program (e.g., see Bendersky, 1978). Such programs are of particular value, since the basic validity of curriculum-referenced instruments lies in the degree to which they reflect the actual curriculum of the programs employing them. If the program changes, therefore, additions, deletions and revisions of items in the checklist are likely to be advisable.

THE VALUE OF CURRICULUM-REFERENCED INSTRUMENTS

Using a curriculum-referenced instrument alleviates many of the problems associated with subject-unique time series data. First, since a well constructed curriculum-referenced instrument will assess progress in both programmed and nonprogrammed areas, it is possible to evaluate a child's progress from a broad perspective--thereby reducing the probability of a "so what?" reaction to demonstrated gains. Second, since the performances of all children in the program are held relative to the same set of objectives, group summaries and analyses are possible. Such summaries can be most useful when making overall program development decisions. Third, since a long-term sequence of skill development is assessed, and since it is possible to analyze progress records across children, certain types of extrapolation procedures can be more easily justified. The problems with extrapolation mentioned earlier centered around the estimation of the time when skill development might reasonably be expected to have begun, possible confounds with age and level of deficit, and problems in determining whether skill development was linear or curvilinear. By measuring a sequence of skills, it is possible to develop a sequence which reaches back to some fairly well defined opportunity for the initiation of skill development. With a preschool scale, for example, the developmental sequence can be extended back to the birth of the child. Linearity of expected development can be assumed if the scale was properly constructed and tested, and even retested longitudinally after the child

enters the program and/or by a cross-sectional analysis of those children already in the program. Possible confounds with age and level of deficit can also be tested cross-sectionally by analyzing the relationship between rate of progress through the curriculum and age and/or level of deficit for children already in the program. In many cases, therefore, it should be possible to estimate meaningful and believable baselines against which further development can be compared (e.g., Haring, 1978; White & Haring, 1978).

ADDITIONAL BENEFITS

As mentioned earlier, no ongoing system of evaluation is likely to survive if it does not provide some immediate benefit to the educators who must make it work. Fortunately, the curriculum-referenced approach offers benefits to teachers and administrators alike:

1. PROGRAM GOALS are more clearly defined by a curriculum-referenced instrument than by almost any other description of program intent conceivable. It is easier, therefore, for funding agencies to determine the program's potential merit, for agencies and parents to determine whether the program might represent a reasonable placement option for a particular child, and for teachers to decide exactly where their instructional emphasis should be placed.

2. GREATER PROGRAM BALANCE can be achieved if the curriculum-referenced instrument is itself well-balanced in all major areas of developmental concern. Teachers need not actually instruct each child in all areas of development; they can concentrate on what they feel are the child's greatest needs. If such an emphasis is appropriate, then the child will continue to develop in all areas (as measured by the curriculum-referenced instrument). If, on the other hand, overemphasis of a subject or skill area is detrimental to the child's continued progress in other areas, that will be easily determined.

3. CONTINUED REFINEMENT OF THE CURRICULUM is facilitated by noting areas in which most children in the program have special difficulties. Sequence and linearity analyses will also help to pinpoint areas in the curriculum which might best be rearranged, restructured, or refined in some way.

4. THE DEVELOPMENT OF INDIVIDUALIZED EDUCATIONAL PROGRAMS (IEPs) is facilitated by providing a ready-made list of the most common objectives which might be chosen for a child. Records of which objectives are most typically selected, which are least often selected, and which are usually modified extensively in some way, will also provide additional information for designing changes in the basic curriculum and the curriculum-referenced instrument itself.

5. THE COMPILATION OF LONGITUDINAL TIME-SERIES DATA which is essentially comparable across children and subprograms provides the basis for the analysis of major changes in the basic program. Of course, the precision of such analyses will depend, in part, on the frequency with which data are collected and the frequency of changes in the program to be analyzed. With many programs, where many of the objectives are based on the expectancies of children in alternative program placements, there is even a relatively high probability that actual preprogram and/or postprogram data might be available for more precise time-series analyses of program impact.

STILL NO PANACEA

Curriculum-referenced approaches to evaluation alleviate some problems, but they are still a long way from being a panacea. They avoid some of the difficulties in applying national norms to special subgroups in the population, but only at the expense of providing more idiosyncratic data of less use in cross-program comparisons. Similarly, they enable the overall progress of children in a given program to be assessed with a common metric, but only by ignoring at least some of the important individual differences between children. Devising objective lists which can be applied to more programs aggravates the problem of accounting for individual differences, and greater specialization to account for individual differences makes the metric less useful for cross-program evaluation. Striking a reasonable balance can be difficult, and the process of validating an instrument once the basic issues of its design have been resolved can be a long and arduous task. Nevertheless, the curriculum-referenced approach

does appear to offer at least a reasonable chance of placing individual time-series data in a broader perspective for overall program evaluation in a way that is useful to educators. Serious consideration should be given to its use and further development as an alternative or adjunct to more traditional group or single-subject program evaluation strategies.

REFERENCES

Barton, E. M., & Ascione, F. R. Sharing in preschool children: Facilitation, stimulus generalization, response generalization, and maintenance. Journal of Applied Behavior Analysis, 1979, 12, 417-430.

Bendersky, M. Uniform performance assessment system: Birth to 6 year level tester's manual, 1968. (Available from College of Education, Experimental Education Unit, Child Development and Mental Retardation Center, WJ-10, University of Washington, Seattle, WA 98195).

Bryk, A. S., & Weisberg, H. I. (Co-principal Investigators). Evaluation methodology for early childhood special education programs. Grant application submitted to the U.S. Office of Education (Handicapped), 10 January 1977, Huron Institute, Cambridge, Massachusetts.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. (Also published as Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.)

Cooley, W. W., & Lohnes, P. R. Evaluation research in education: Theory, principles and practices. New York: John Wiley and Sons, 1976.

English, H. B., & English, A. C. A comprehensive dictionary of psychological and psychoanalytical terms. New York: David McKay Co., 1958.

Haring, N. G. Center for Severely Handicapped: Report to the Joint Dissemination Review Panel. Unpublished manuscript, 1978. (Available from Experimental Education Unit, University of Washington, Seattle, WA 98195).

Hersen, M., & Barlow, D. H. Single case experimental designs: Strategies for studying behavior change. New York: Pergamon Press, 1976.

Horst, D. P., Tallmadge, K., & Wood, C. A practical guide to measuring project impact on student achievement. Mountain View, CA: RMC Research Corporation, 1975.

Kratochwill, T. R. Single subject research: Strategies for evaluating change. San Francisco: Academic Press, 1978.

Lahey, B., Gendrich, J., Gendrich, S., Schnelle, J., Gant, D., & McNees, M. An evaluation of daily report cards with minimal teacher and parent contacts as an efficient method of classroom intervention. Behavior Modification, 1977, 1, 381-394.

Ragland, E., Kerr, M., & Strain, P. Behavior of withdrawn autistic children: Effects of peer social initiations. Behavior Modification, 1978, 2, 565-578.

Sailor, W., & Mix, B. J. TARC assessment system. Lawrence, KS: H & H Enterprises, 1975.

Sidman, M. Tactics of scientific research: Evaluating experimental data in psychology. New York: Basic Books, 1960.

Smith, D., & Lovitt, T. The use of modeling techniques to influence the acquisition of computational arithmetic skills in learning-disabled children. In E. Ramp & G. Semb (Eds.), Behavior analysis: Areas of research and application. New Jersey: Prentice-Hall, 1975.

Tallmadge, G. K. The Joint Dissemination Review Panel ideabook. Mountain View, CA: RMC Research Corporation, 1977.

Tallmadge, G. K., & Horst, D. P. A procedural guide for validating achievement gains in educational projects. Mountain View, CA: RMC Research Corporation, 1976.

Walls, R., Werner, T., Bacon, A., & Zane, T. Behavior checklists. In J. Cone & R. Hawkins (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, 1977.

White, O. Adaptive performance objectives: Form vs. function. In W. Sailor, B. Wilcox, & L. Brown (Eds.), Instructional design for the severely handicapped. Baltimore: Brookes Publishing, 1980.

White, O., & Haring, N. Evaluating educational programs serving the severely and profoundly handicapped. In N. Haring & D. Bricker (Eds.), Teaching the severely handicapped (Vol. 3). Seattle: American Association for the Education of the Severely and Profoundly Handicapped, 1978.

White, O. R., Reavis, K., & Kunzelmann, H. Child performance based decisions for screening, placement, and adaptive program planning. Presentation at CEC Topical Conference on Individualized Educational Plans, Albuquerque, New Mexico, February, 1978.

Winer, B. J. Statistical principles in experimental design. San Francisco: McGraw-Hill, 1962.

Measuring Child Progress: Large Group Design and Norm-Referenced Alternatives

Robert Sheehan

I welcome you again to the working sessions of this topical conference on evaluation. We now begin the difficult task of analyzing the several individual areas that must be evaluated in early childhood special education programs and responding to the project-specific concerns that many of you have in these areas.

I will be addressing the area of measuring or documenting child progress, focusing primarily on the use of large group and norm-referenced alternatives. I hope that it will become obvious very quickly, however, that your choice of design (such as norm-referenced versus single subject) is definitely not the first question to be considered. My reasons for this qualification are an important part of my address to you today.

As an introduction, I would like to point out several perspectives which underlie my current approach to the documentation of children's progress. I will introduce the large group and norm-referenced evaluation designs that are currently in use in the field of early childhood special education. I will also briefly discuss the instruments currently used in the field. This discussion of designs and instruments will be followed by a decision making matrix which, though only in the formative stages of development, will be of use in operationalizing your roles as decision makers in your program evaluation efforts.

UNDERLYING PERSPECTIVES

This address is part of a sequence of discussions regarding the documentation of handicapped and non-handicapped children's progress in intervention programs. As it has heavily influenced my thinking about the evaluation efforts of the field, an outline of this sequence may be useful to you.

My first attempt to discuss the activities of the field in the documentation of progress consisted of a very long review paper which was written with Dr. Barbara Keogh of UCLA this past summer (Sheehan & Keogh, 1978). This paper convinced us that there is much more to know about the use of evaluation designs, instruments, and analytic techniques to document the progress of children in intervention programs. My presentation today about evaluation designs is, in a large part, drawn from this review paper.

I was then asked by TADS to address a conference in Nashville similar to this. The first topic I addressed dealt with the use of instruments in the documentation of mild/moderately handicapped children of preschool age (Sheehan, 1978a). The second address pertained to the implications of evaluation design on instrument selection in early childhood special education (Sheehan, 1978b). The outcome of these two efforts convinced me that one cannot separate the issue of methodology from programmatic concerns. It is appropriate that Owen White and I are both talking with you this morning, because any choice of evaluation designs must be made from a preliminary vantage point which accepts the usefulness of both large and small group designs, depending upon the programmatic questions that you are seeking to answer with your documentation efforts.

As a result of dealing with the documentation of children's progress on both theoretical and practical levels, I have developed three perspectives on the field which, though definitely interrelated, are useful to discuss individually in a particular sequence. The first of these is that there is a strong interdependence between your evaluation design and the instruments that you select to document the progress of young handicapped children.

The second perspective, that follows from a recognition of methodological interrelationships, is that none of the decisions about designs and instruments can be made without an additional decision pertaining to the type of argument you wish to make with your documentation efforts regarding children's growth and development. This decision must be made with reference to the handicapping condition or "at-risk status" of the children that you serve.

The third perspective has been mentioned before but deserves repetition at this point. Considering the interrelated nature of the decisions which must be made in your documentation or progress efforts, you must move quickly to a view of yourselves as decision makers. You are the ones who must ultimately coordinate the decision making process, insure that the interrelations between methodological concerns are recognized, and insure that your programmatic concerns (pertaining to the type of arguments you wish to make) will be adequately dealt with by the methodological decisions.

These three perspectives form the basis for the rest of this address. They have guided my review efforts and have also been used to generate a decision making matrix for early childhood special educators. I will share this matrix with you following a discussion of designs currently in use.

DESIGNS CURRENTLY IN USE

I am sure that many of you have taken an introductory research methods course at some point in your college careers. If so, you may be familiar with the sixteen research designs of Campbell and Stanley (1966) which are reviewed in terms of their internal and external validity. Of these sixteen designs, three are currently in use in the field of early childhood special education. Two other types of designs currently being used in the field of early childhood special education are not discussed by Campbell and Stanley, and thus are new or significant variations of the traditional designs.

The Campbell and Stanley (1966) approach to the discussion of these designs is to critique their validity independent of the situation in which the designs are used. This has methodological purity but may in fact

result in the abandonment of adequate designs for certain types of situations. To present another approach to a critique of designs currently in use in early childhood special education, I would remind you of one of the basic principles of any research methods class--namely, that research design is utilized to test and rule out alternative hypotheses. The question, as I consider the five types of large group and norm-referenced designs, is whether or not the use of each design for certain types of arguments in the field leaves unresolved rival hypotheses or alternative explanations for the resulting data.

The three research designs proposed by Campbell and Stanley (1966) which as currently being used to document the progress of handicapped children are: (1) the one group pretest-posttest design; (2) the time series design; and (3) the non-equivalent control group design. In addition, two other designs being used are: (4) the one-shot prediction design and (5) the pretest-posttest prediction design.

THE ONE GROUP PRETEST-POSTTEST DESIGN

The one group pretest-postest design in early childhood special education involves two assessments of handicapped children and the presentation of these data as documentation of child progress. These assessments most logically are made prior to and following intervention. The critical features of this design are that two assessments are made of handicapped children and that these assessments only reflect children who have received the educational intervention.

The one group pretest-posttest design has been criticized by Campbell and Stanley (1966) as a bad example of research design. They point out that, without a comparison, evidence of change from pretest to posttest could result from many factors including but not limited to the effects of intervention.

Campbell and Stanley's logic is certainly sound in most educational settings. However, we are not convinced that this design is totally lacking in scientific merit in all cases. Our first hesitation is a very pragmatic one. P.L. 94-142 now requires that all handicapped children be served by states and local educational agencies (LEAs). As this law becomes fully operational, the possibility for any control group or

comparison group decreases. It may be that in many cases early childhood special education programs will not be able to find accurate comparisons and will fall back on this type of design by necessity. This is not merely a problem of inability to randomize (which Campbell and Stanley respond to with quasi-experimental designs). It is a more fundamental problem of absence of comparable children.

Our second hesitation to immediately criticize this design is more specific to the nature of some children's handicapping conditions. Consider for example the frequently occurring case of an early childhood special education project which serves profoundly handicapped children. These are children whom our best clinical judgment indicates will be totally dysfunctional by age eight. If a program can demonstrate that children who have been appropriately diagnosed as profoundly handicapped have shown developmental progress, or acquired skills where before there were none, these data should not be ignored nor dismissed.

The underlying point to this argument is that, while there is no formal control or comparison group, there are expectations for these children (clinically-derived) which have been exceeded. For these children, demonstrating any response to treatment is a positive indicator of success.

However, there are very definite qualifications to the use of this design in early childhood special education. The first qualification is that there must be an absence of any set of data which can be used as a comparison or control for the data describing children who have had intervention. The second qualification to the use of this design is that there must be some clinically-derived concept of how children would fare in the absence of intervention and what responses are likely from intervention. This qualification has some very problemmatic implications which suggest that this documentation design will not be effective for very long. If the design yields evidence which exceeds the clinical expectation, then the clinical expectation should be replaced or revised and more accurate predictions with an empirical basis should be developed.

An example of revising clinical judgments may be found with Down's syndrome children. Pediatricians have long argued that Down's syndrome children cannot be taught. Parents have been advised to institutionalize

these children as early as possible. Recent evidence indicates that these clinical impressions are quite inaccurate and that Down's syndrome children can be taught. The conclusion to be drawn is that, as our data base regarding the growth of handicapped children expands, the usefulness of this design declines.

An example of this type of design can be found in the Joint Dissemination Review Panel (JDRP) submission of the Portage Project. This project presents data that multiply handicapped children showed a statistically significant increase on the Alpern-Boll Developmental Inventory and the Stanford-Binet Test. Whether or not these children would have shown a statistically significant increase without the intervention is unknown.

A second example of this design is found in the JDRP submission of the Rutland Center. It reports that, for children who were emotionally disturbed, teachers felt that behavior problems decreased: "...teachers made pre/post ratings for 15 children. The results of the analysis indicated that the percent of 'high priority' problems perceived by teachers decreased from the time of referral to the time of termination for 49 of the 63 problem items" (p. 6). The strength of this argument lies in the relationship between the handicapping condition of the child (emotionally disturbed) and the reviewers' acceptance of the 63 problem items as important for the educational setting.

There are other examples where the usefulness of such a design might be questioned. Pefley and Smith (1976) report data from a project serving a wide range of handicapped children in which pretest and posttest scores are acquired on the Bayley Scale and the Denver Developmental Profile. The only children who were tested were those who received intervention, and the argument for child progress is made simply on the change of scores from pre- to posttest (p. 66-68).

Yet another example of this type of documentation can be found in the report of another project (Pefley & Smith, 1976). The change in children's adaptive behavior during a six-month period of time was assessed. The average pre and post scores for the 58 children were reported as documentation of child progress. The children were simply identified as "handicapped" (p. 69-70).

The difficulty which makes the use of this design so unacceptable in the two final reports just mentioned is

that not enough evidence is presented which describes the handicapping conditions of the children being considered. The reader has no knowledge of the kinds of children and their clinical expectations.

Reviews of the current activities of early childhood special education projects and discussions with professionals in the field indicate that the one group pretest-posttest design has been used frequently in the past, and in fact it has been used too much. Too many arguments for child progress have been made with this type of design. Despite this problem, there are and will continue to be instances in the future when our knowledge of what to expect from some handicapped children is so inadequate that any progress is a good sign. This factor should be considered in reviewing the documentation designs of early childhood special education projects.

TIME SERIES DESIGN

In the time series design, a number of assessments are made of handicapped children and these assessments are presented as evidence of child progress. The critical features of this design are that the assessments are made frequently (for example, daily, weekly, monthly), and the assessments are made only on children receiving the educational intervention. The time series design which is actually used in early childhood special education is a variation on Campbell & Stanley's (1966) design in which repeated measurements are taken before and after a treatment but not during a treatment. The design used in early childhood special education involves measurement before, during, and sometimes after an educational intervention. Huck, Cormier, & Bounds (1974) identify this variation of a time series design as a single group with continuous single treatment and possibly withdrawal of treatment.

As indicated by Campbell & Stanley (1966), the time series design is stronger than the one group pretest-posttest design. It has fewer sources of invalidity. One of its sources of invalidity is the problem of history. Something could happen during the course of intervention which is unrelated to the intervention and might cause a

change in children's progress. In most cases the possibility of this happening is quite unlikely and would be noticed if it did occur (for example the development of a miracle drug that made handicapped children like nonhandicapped children).

This design is potentially a very powerful one in the field of early childhood special education. The repeated measurement of child progress insures that children's performance will be closely and frequently examined. With each assessment, more information is gained about the progress of children. Once again, though, there is an implication of this design which is not immediately apparent. Following the first two or three assessments, it is very likely that a body of data will be developed which represents the current documentation of child progress. If this is the case, early childhood special education projects can build these data into their design as a point of comparison and the design will shift from a single group time series design to a multiple time series design with a comparison group (in this case the comparison group is the constantly changing set of baseline data).

An example of this type of design was found in a review of the submissions of early childhood special education projects to the Joint Dissemination Review Panel. The first submission was that of the Central Institute for the Deaf, which measured the growth in language ability of deaf children over a 2½-year time span with five different assessment periods. This project reported that scores increased consistently and reliably throughout all age ranges of children. The data were presented graphically, and trend analysis was performed on the profile of scores.

The time series design seems to have the most potential to early childhood special education projects where data are of use to project personnel for program planning (as in the Teaching Research program) or where children will be in the program for quite a long time, permitting several six-month assessments (as in the Central Institute for the Deaf program). Programs which have children for shorter periods of time and who do not use formal child documentation data for program planning seem less likely to use this design.

THE NON-EQUIVALENT CONTROL GROUP DESIGN

The non-equivalent control group design is used quite frequently in general educational research, however it is not found frequently in early childhood special education programs. The design consists of two assessments of two groups of children. The assessments are usually made prior to intervention (pretest) and following intervention (posttest). The two groups of children include those receiving the educational intervention and others who are judged to be similar to those receiving treatment. No random assignment to groups occurs. As indicated earlier, one reason why this design may not be used frequently is the lack of suitable control or comparison children. With the impact of P.L. 94-142 being felt throughout the country, the likelihood of obtaining data on a group of handicapped children who have not received intervention dwindles. This leaves early childhood special education projects with several alternatives. They can continue to try to locate a comparable group of children somewhere in the country. The catch to this approach is that if these children are located and are not being served, this fact should be reported and the children should be served.

A second alternative available to early childhood special education projects is to locate children who are not handicapped and attempt to equate them either through matching subject characteristics or covarying existing differences. This requires a sophistication that may not be present in early childhood special education programs and the success of this approach is in no way assured.

Yet a third alternative is to recognize the "model" status of First Chance programs and to consider as a viable comparison group children who are receiving treatment in other intervention programs. This alternative poses the greatest potential considering the need for cost-efficient intervention and the testing of "models." It also recognizes that professionals in the field no longer need to be told that "treatment" is better than "no-treatment" but instead, "What works best?"

An example of the use of the non-equivalent control group design can be found in the JDRP submission of Project ERIN (1978). In this project, child progress data were collected on one group of children (N=42) who

received early childhood special education intervention. Data were also collected on a group of children (N=32) who did not receive the intervention. Children were matched on the basis of average pretest scores, and both experimental and comparison children were tested with three standardized instruments. Children were tested at the beginning and end of the six-month period and their scores were compared. The comparison children were not denied intervention but were selected from other sources on the basis of matching variables.

This design has not been used widely. It is interesting to note that of the 27 early childhood special education projects reviewed in It's Monday Morning (Pefley & Smith, 1976), not one project had identified comparison groups. It is quite likely that in view of recent legislation and the pressures at work in this society to educate all handicapped children, this design in its treatment vs. no treatment form will not be used extensively in early childhood special education. Instead, the field may accept the viability of a Treatment A vs. Treatment B comparison within this design.

THE ONE-SHOT PREDICTION DESIGN

In some form or another, prediction designs are used most frequently by the vast majority of early childhood special education projects. Because of this heavy reliance on prediction designs, careful attention should be paid to their underlying assumptions.

The one-shot prediction design has two critical features. The first is that this design involves only one assessment of handicapped children's progress. The second is that data are collected only on children receiving educational intervention and then are compared to a hypothetical comparison group which might be derived in several ways. In essence, comparison in the prediction design is made to an estimate of what children's progress would have been in the absence of intervention.

The hypothetical comparison group is computed in any one of several ways. The simplest example of this occurs when early childhood special education projects argue that children who are chronologically five years of age are functioning at developmental ages of five years. These developmental ages are derived by considering the

average performance of the five-year-olds who were involved in the standardization of the test or instrument used to compute the developmental age of the children in the project. What the early childhood special education project has done (perhaps unknowingly) is compared their data to norming data.

This particular documentation approach is not totally new to education. For years, grade school teachers have examined the percentile performance of their children at the beginning and end of the year. With the ready availability of instruments in early childhood special education which yield developmental ages, standard scores, percentile rankings, and so forth, the stage is set for this type of argument.

In fact, very few early childhood special education projects utilize the one-shot prediction design. They use instead the pretest-posttest prediction group design. The assumptions underlying these two designs are slightly different (as the latter design refers to developmental growth). For the present, the one-shot prediction design is making one major assumption: that the children in the norming population are accurate and reasonable comparisons for the handicapped children who have received the intervention. This prediction might be very appropriate for children who are only mildly developmentally delayed. The goal for these children should perhaps be normal development at some point in time. The assumption may be far less appropriate for children who are multiply handicapped, deaf, or blind. Normal development may not be an appropriate point of comparison for these children.

An example of this type of documentation design is provided by the JDRP submission of the Down's syndrome program at the University of Washington. This report argues that "analysis of performance on the Denver Developmental Screening Test administered to 50 children across the four Down's syndrome programs revealed that a high percentage of the items on the test were passed in August, 1974 (100 percent of the items passed represents normal performance)" (p. 2). The assumption underlying this argument is that the average performance of the children on whom the Denver was normed is an appropriate comparison.

This design has several immediate advantages for early childhood special educators. First, data need only

be collected on the children who are receiving the intervention. This spares project personnel from having to deny services to children or engage in extensive matching and covarying procedures with a limited likelihood of success. The second advantage of this approach is that it suggests a goal with which staff, parents, and funding agencies can identify: normal development.

However, there are some real disadvantages to this approach. First, a one-shot assessment may not be very accurate at all. Results of such testing might be very ambiguous or misleading. A second problem is that normal development may be much too conservative a standard for comparison. A program working with multiply handicapped children would be setting itself up for failure if it continually aspired to achieve normal developmental status by a certain point in time. These disadvantages have resulted in early childhood special educators turning in increasing numbers to the next type of design.

PRETEST-POSTTEST PREDICTION DESIGN

The pretest-posttest prediction design is different than the one-shot prediction design in that two assessments of handicapped children's performance are made. These assessments usually occur prior to intervention (pretest) and following intervention. A second critical feature of the design is that handicapped children's performance (or in many cases developmental growth from pretest to posttest) is compared to an estimate of that performance in the absence of intervention. As in the one-shot prediction design, the estimate of what children's progress would have been in the absence of treatment is derived in several ways.

Many early childhood special education projects use the pretest-posttest prediction design. Perhaps more important than the use of this design is the fact that the national evaluation of the Handicapped Children's Early Education Program was conducted using two types of this documentation design. One can only speculate as to the influence of this evaluation on early childhood special education projects, but it is likely that these projects have been encouraged in their use of the pretest-posttest prediction design by the national evaluation.

The national evaluation of early childhood special education projects funded by the Bureau of Education for the Handicapped (BEH) was conducted by Battelle Institute (Stock, Newborg, Wrek, Schenck, Gabel, Spurgeon, & Ray, 1976). The Battelle evaluators, using an instrument of child progress (CEEDI) which they developed, devised two methods to compute the predicted control group. The first involved the use of age-matched subjects. Each child was matched at posttest age with another child who had not received intervention. Thus comparisons were made of the posttest scores of one group with essentially what were pretest scores of the matched group. More specifically, matched pairs of children were devised by the Battelle staff: "...the first member of each pair being of the same age at pretesting. The mean posttest score of all the first members of each pair represents the CEEDI test performance for children having had HCEEP project experience. The mean pretest scores of all the second members of each pair are taken as an estimate of CEEDI test performance for handicapped children of a comparable age in the absence of project experiences. Any difference between the mean of these two groups, then, is indicative of program impact" (p. 66).

As all the children assessed in the Battelle evaluation were experiencing their first year in those particular early childhood special education programs, the evaluations assumed that an entering child of five years of age, for example, would be a good comparison for an exiting child of five years of age.

Rather than rely exclusively on the matching method, the Battelle staff also utilized another method of computing a predicted control group. This method was statistical yet less complex, as no matching was required. The evaluators pretested all children in the early childhood special education programs with the CEEDI. They then computed the magnitude of relationships between the children's pretest score and age. This linear equation was used to predict a posttest score based on the age increase of children while they were in the program. In this method the standard of comparison is the predicted growth level based on the rate of growth at the time of pretest.

In one form or another, the pretest-posttest prediction design is the design of choice of many current early childhood special education programs. An example

of the age-score relationship for control group prediction may be found in the utilization of developmental quotients. The Developmental Profile is an instrument which yields a developmental age in each of five areas. A common practice is to take the pretest developmental age in a given area (DA_1) and divide by a child's chronological age at pretesting (CA_1). This yields what is commonly called a developmental quotient (DQ). This figure is interpreted to be an indication of the "rate of growth" which a child is exhibiting at pretest time.

The developmental quotient, or rate of growth, can then be utilized in a statistical formula to predict where a child would be functioning at posttest time if the rate of growth remained constant. To do this the developmental quotient (DQ) is multiplied by the child's age at posttest time (CA_2). The product is a predicted developmental age at posttesting. The children's actual developmental age at posttesting (DA_2) is then compared to the predicted score and an estimate of program impact is obtained from the difference between these scores. This method can be expressed in the form of the following equation:

$$\text{Program impact} = (DA_2) - CA_2 \cdot \frac{DA_1}{CA_1}$$

A similar approach is found in the Research Management Corporation's (RMC) development of three evaluation models for federally funded Title I programs. These norm-referenced models utilize the percentile performance of a group at pretest time to predict the percentile performance at posttest time. The difference between these predictions is considered to be evidence of program impact (Tallmadge, 1977). The same basic principles are at work in the designs proposed by RMC and those utilized by Battelle Institute. Both assume a constant rate of growth over the period of intervention, thus allowing the computation of given predicted outcomes or developmental accomplishments. Where outcomes are discrepant from prediction (i.e., where the rate of growth is accelerated or decelerated relative to the predicted rate), program influence is assumed.

One major difference between the Title I models and the prediction models used in early childhood special education is that the likelihood of finding disadvantaged

children in the norming populations of the standard instruments for Title I evaluations (e.g., MAT or CAT) is fairly high. On the other hand, the likelihood of finding handicapped children in the norming populations of tests commonly used in early childhood is extremely low. As a result, such assumptions as constant rate of growth or acceptability of norm performance might be seriously violated in early childhood special education programs.

An example of an early childhood special education project using the pretest-posttest prediction design can be found in the JDRP submission of the Portage Project. In this report evaluators argued: "Results of the evaluation of child progress for 1974 indicated that the children in the project gained an average of 15 months in an eighth-month period of time as measured by the Cattell Infant Test and the Stanford-Binet Intelligence Test. Because the average IQ of the target population is 75, it would be expected that on the average, the normal rate of growth would be 75 percent of that of the child with normal intelligence; therefore, using mental ages, expected average gains would be about six months in an average eight-month period of time" (p. 5).

This design for documenting child progress--constructing a predicted or hypothetical control based on the entry characteristics of program children--is being used by early childhood special education projects with such diverse instruments as intelligence tests, achievement tests, and developmental profiles. Its advantage is, of course, that ethical concerns are set aside: no children are denied services in order to maintain treatment-control conditions.

INSTRUMENTS CURRENTLY IN USE

As I indicated earlier, one of my underlying perspectives is that decisions about designs are interrelated with decisions about instruments. Given this assumption, I would like to briefly discuss the instrumentation being used in the field and some of the issues surrounding the use of these instruments.

There has been a rapid proliferation of instruments in the field of early childhood special education during the last five years. At UCLA we conducted a review of instruments being used to assess the performance of

infants. Our review uncovered approximately 300 instruments, and we know that there are more.

Cross and Johnston (1977) identified 98 instruments which were being used by programs in the Handicapped Children's Early Education Program (HCEEP). This review indicates that, as extensive as it is, several current instruments are not identified.

Recently, a review has been conducted of 27 early childhood special education projects funded by the Bureau of Education for the Handicapped (BEH) in July, 1972. This review indicated that 27 projects employed 60 different assessment instruments in some aspect of their program. In addition, 15 of the projects developed their own assessment instruments. The ways in which the 60 instruments were used in the 27 early childhood special education programs is not specified. Presumably this usage would include identification, diagnosis, assessment, and evaluation purposes.

Another fact which is not known is why 15 of the projects developed their own instruments. Our own experiences with early childhood special education programs indicates that many projects develop their own curricula and assessment instruments so that they can be cross referenced and used in an integrated fashion with the specific type(s) of children being served by those projects. One research effort of merit is an exploration of why early childhood special education projects develop their own instruments.

The names of the instruments commonly used by the 27 programs and the frequency of their usage are reported in Table 1.

At this point, one very real problem with instruments must be pointed out. The instruments currently used in the field were not designed for documentation purposes. They were designed either to identify children as handicapped, to conduct differential diagnosis, or more recently to facilitate curricular decisions. They were not designed to assess the changing nature of children's performance. Items in these instruments were selected because of their high frequency of occurrence at predictable ages. Due to this difficulty, consider the results of your instrumentation to be very conservative.

TABLE 1
Frequently Used Instruments

Instrument	No. of Projects
Denver Developmental Screening Test	4
Illinois Test of Psycholinguistic Abilities	3
Peabody Picture Vocabulary Test	6
Portage Project Checklist	4
Bayley Scales of Infant Development	7
Alpern-Boll Developmental Profile	6
Stanford-Binet	5
Frostig	3

Source: Pefley & Smith, 1976.

A DECISION MAKING MATRIX

Over the period of the last six months I have been considering the possibility of a decision making matrix which early childhood special educators could use to focus their search for instruments and designs. It is clear to me that such a matrix must be responsive to several dimensions or concerns. One obvious concern, highlighted by Campbell and Stanley (1966), is for the methodological adequacy of the evaluation designs, and it often centers around the types of comparisons being made with the particular design being used. This comparison answers the question: "What is it that the progress of children receiving the intervention is being compared to?" In our review of the field, we found at least four levels along which this question could or did vary: (1) no comparison; (2) a randomly assigned control group; (3) a quasi-control group; (4) a statistical or predicted control group.

A second concern to which a decision making matrix must be responsive pertains to the frequency of assessment. How often during the course of a year or a treatment period are child progress data to be collected? We found answers to this question occurring at three levels: (1) single assessment; (2) two assessments; (3) multiple assessments. It is clear that the answer to this

question affects both the selection of evaluation designs and instruments.

There is a third concern to which a decision making matrix in early childhood special education must be responsive as well. This concern takes the form of the question: "Relative to the type(s) of children you are serving, what type of child progress argument do you want to make with your evaluation?" I have only begun to consider the possible answers to this question. It seems that at least four answers might be made, which are actually the child progress goals that you have for all or subgroups of your children. The answers include: (1) normal development in all areas; (2) normal development in several areas; (3) any development in all areas; (4) any development in several areas.

Let me elaborate on the ways in which the answers to this third concern relate to the children that you are serving. If you are working with children whose handicapping condition consists of an overall developmental delay, you may have as a very realistic goal the attainment of normal development for these children in all areas. If so, your instrument selection should assess overall developmental progress and you might be able to use a norm-referenced strategy such as a prediction design to compare your children's progress to normal children.

On the other hand, you may be working with learning disabled children who function normally in most areas but exhibit an uneven profile. If this is the case, your goal may be normal development in one or several areas (those which are in greatest need). Your choice of instruments would then focus on specific developmental areas or subtests and you may utilize a design which makes reference to a norm in those particular developmental areas.

In your early childhood special education programs you may be serving children whose handicapping conditions are quite involved (for example, multiply handicapped children). With these children, you might want to demonstrate any progress at all in the beginning of your intervention efforts and then adjust those expectations as the intervention continues. In this case, your selection of instruments would likely reflect the acquisition of specific behavioral responses.

A fourth possibility would be the type of argument that you wish to make for children who have specific handicapping conditions in relatively few areas. For example, blind children would always differ from sighted children on visual performance tasks but there may be specific behavioral accomplishments that you would want to focus on. In this case, an instrument must be selected which reflects the acquisition of these skills and an evaluation design would be chosen which highlights the children's performance relative to their earlier nonperformance on these items.

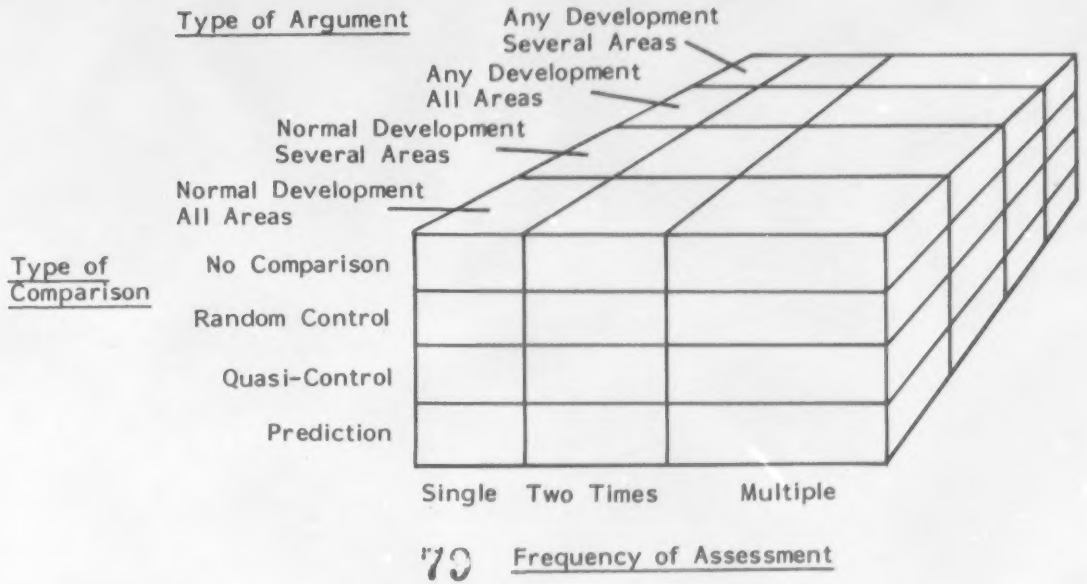
The question now becomes, "How do we translate such concerns or dimensions into a matrix and how do we use this matrix for decision making?" I have developed a three-dimensional matrix (Figure 1) which captures the dimensions or concerns which I have discussed and also identifies the levels of possible responses to these concerns. I should quickly point out that you may be able to expand this matrix by adding additional dimensions or levels.

The matrix can be used in a decision making fashion by simultaneously addressing the following three questions:

1. Considering the children you are working with, what type of argument do you want to make with your documentation efforts regarding children's growth and development?
2. What type of comparison can you (must you) make in your evaluation design to insure that your results do not occur by chance?
3. How frequently can you (must you) gather data on children's progress?

It seems to me that use of such a systematic process would guide our decision making in early childhood special education. To demonstrate this, let's take an example of a program that is working with 10 children who are language delayed. We would first decide that the major argument which we wish to make is that, as a result of our program, the children have moved toward normal development in their expressive and receptive language. This would then lead you to selection of a category of instruments dealing with these language areas.

FIGURE 1
Decision Making in Documenting Child Progress
in Early Childhood Special Education



Following the identification of this category of instruments, you might then consider your possible comparisons. If you cannot locate a control or comparison group, you are left with some type of norm-referenced or prediction argument. You cannot select the "no comparison" option because your children will progress in their language abilities in some fashion anyway.

The selection of a norm-referenced strategy then suggests that you must locate an instrument which has appropriate norms or can yield valid expectations. This further narrows your possible choice of instruments.

This decision making process may not yield completely specific answers. However, by using such a process, you can identify when to turn to experts in the field. For example, an expert can help locate an instrument when your requirements for that instrument are known. Similarly, an expert can help suggest a design when your instrumentation needs are clear and when you have made decisions on the availability of appropriate comparison groups.

SUMMARY

It is clear that the field of early childhood special education is rapidly becoming more concerned with the need to measure the progress of handicapped and nonhandicapped children. This concern has led to the consideration of several new evaluation designs and the development of many new instruments to assess child progress. Early childhood special educators must adopt a view of themselves as decision makers. The decision making process must include the following considerations:

1. There is a strong interdependence between the evaluation designs that you select and the instruments which you use to measure children's progress.
2. Evaluation decisions of design and instrumentation must be made with reference to the type of argument that you wish to make. This reflects the types of children that you are serving.
3. Early childhood special educators are decision makers.

4. There are at least five types of evaluation designs in use in the field today. They include: one group pretest-posttest design; time series design; non-equivalent control group design; one-shot prediction design; pretest-posttest prediction design.
5. There is an extremely large number of instruments in the field today. Not many instruments are used by most programs. The instruments in the field today were not designed for documentation of progress and may not be responsive to the changing child.
6. Three major questions must be addressed by early childhood special educators in their decision making efforts. These include:
   a. Considering the children you work with, what type of argument do you want to make regarding children's growth and development?
   b. What type of comparison can you (must you) make in your evaluation design to insure that your results do not occur by chance?
   c. How frequently can you (must you) gather child progress data?

REFERENCES

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Cross, L., & Johnston, S. A bibliography of instruments. In L. Cross & K. Goin (Eds.), Identifying handicapped children: A guide to case finding, screening, diagnosis, assessment and evaluation. New York: Walker Publishing Co., 1977.

Huck, S. W., Cormier, W. H., & Bounds, W. G. Reading statistics and research. New York: Harper and Row, 1974.

Pefley, D., & Smith, H. It's Monday morning. Chapel Hill, NC: Technical Assistance Development System (TADS), 1976.

Sheehan, R. Measuring progress in the mild/moderately handicapped child--preschool age. Paper presented at TADS Topical Conference on Evaluation, Nashville, TN, Fall 1978a.

Sheehan, R. The implications of evaluation design on instrument selection. Paper presented at TADS Topical Conference on Evaluation, Nashville, TN, Fall 1978b.

Sheehan R., & Keogh, B. Documenting child progress in early childhood special education: The state of the art (draft). Los Angeles: Project REACH, UCLA, 1978.

Stock, J. R., Newborg, J., Wrek, L. L., Schenck, E. A., Gabel, J. R., Spurgeon, M. S., & Ray, H. W. Evaluation of Handicapped Children's Early Education Program (HCEEP). Final Report to Bureau of Education for the Handicapped, U.S. Office of Education, 1976. (Contract #OEC-0-74-0402)

Tallmadge, G. K. The Joint Dissemination Review Panel ideabook. Washington, DC: National Institute of Education, & U.S. Office of Education, 1977.

Evaluating Parent Training Programs: Some Considerations

Meave V. Stevens and Roger Kroth

The need for parent training has been widely discussed in the educational literature (Gordon, 1972; Kroth, 1975; Kroth & Simpson, 1977; Radin, 1969; Schaefer, 1972). This new impetus to involve parents in the education of their children is a multidirectional one. Professionals in special education, child development, mental health, and community services are currently involved in advocating and providing training and services to parents. Recent federal legislation requires that parents of exceptional children be involved in certain aspects of their child's education. Books and packaged programs on parenting, offered publicly, have been favorably received and are increasing in number. Many more parents are receiving training and becoming involved in the education of their children. This increased involvement will have an effect on the lives of millions of children, and it is essential that this effect be a positive one (Kroth & Simpson, 1977).

PROBLEMS ENCOUNTERED

Past efforts to evaluate the effectiveness of parent training programs suggest many technical and methodological problems. O'Dell (1974), in a critical review of the literature on training parents in behavior modification techniques, stresses the need for empirical testing of parent programs. O'Dell points out the tendency of researchers to focus on a set of techniques rather than parent change. He states:

> The major problem is the lack of data on producing and maintaining changes in parents. Effective parent training requires three steps. The parent must acquire the modification skills and changes in their own behavior, changes must be implemented with the child, and changes must generalize and persist. Of these three steps, only the implementation phase with the child has received sufficient attention. The tendency to focus on the child has resulted in a neglect of obtaining data on changes in parent behavior (p. 430).

Schaefer (1972) presents a comprehensive review of the early childhood and early education research involving parents as educators. This review strongly supports the efficacy of involving trained parents in the education of their young children. While certain parent training programs are effective for short periods of time, whether or not "brief parent training programs will be sufficient to have long term effects upon parents' education of their children, is as yet undetermined" (Schaefer, 1972).

There is ample and credible evidence in the literature to support parent involvement, in one form or another, as a key factor in increasing child performance (Duncan & Fitzgerald, 1969; Radin, 1969; Schaefer, 1972; Shearer & Shearer, 1972). This evidence, however, is primarily based on child change data or non-experimental assessment. Which parent training programs are most effective, with what population and for how long, is largely undetermined, although a recent review of parent training manuals by Bernal and North (1978), shows some effort in this direction.

The problems encountered in the evaluation of parent training programs are the same as those involved in the evaluation of any social action program. The nature of parent training programs is such that one evaluation procedure may not suffice within a given program or across programs. Evaluations are made for the purpose of providing information for decision making (Stufflebeam, 1971). Decision making regarding parent training programs may require specific information on a number of facets within each program as well as general information on the effectiveness of the overall program.

WAYS TO EVALUATE

Typically, evaluation of parent training programs involves surveying the consumers. Parents are asked to fill out a questionnaire rating the program, trainer and their own progress. Occasionally, teachers may fill out questionnaires, testifying to the effectiveness of the parent program based on observed behavior changes in the child and/or parent. A few commercially available training programs include specific criterion measures of desired outcomes (Wagonseller, 1977). These measures are used following training sessions and serve as short term assessment checks of knowledge or skills gained. Other post measures such as attitude scales or child observation data may be used to determine the presence or absence of a desired outcome.

Perhaps the best evaluation procedure currently used consists of a combination of the above two techniques with the addition of a quasi-experimental, pre/posttest feature. The Parent Involvement Center in Albuquerque, New Mexico, a model demonstration program funded by BEH, is using this combination in its initial evaluation of the effects of available parent training programs. The Appendix contains a parent involvement measure developed at the Parent Involvement Center.

CONSIDERATIONS

While the efficacy of experimental design as a technique for evaluation of broad-aim social programs has been seriously questioned by Weiss and Rein (1970) as well as other research methodologists, it is the writers' opinion that professional as well as practical considerations require the use of experimental methods. At the same time, it is agreed by the majority of research methodologists that not all programs or all facets of program functioning are amenable to experimental or even quasi-experimental evaluations (Campbell, 1970; Smith, 1979).

A useful distinction between demonstration programs and operational programs was made by Suchman (1972). While experimental evaluation was considered useful with

demonstration programs, it was not suggested for use with service (operational) programs. Suchman identified three types of demonstration programs and discussed the use of experimental evaluation with each of the three stages. The first stage, the pilot program, is described as a trial and error phase where staff is involved in testing different structures, methods and approaches. Flexibility is of maximum importance and experimental evaluation is of little use during this stage.

The second stage is the model program. The model program type incorporates the information and experiences gained in the pilot program. Suchman stresses the need for a high degree of organizational structure to insure the success of the model program. This organizational structure as well as the incorporative nature of the model program lends itself readily to experimental evaluation.

The third stage of a demonstration program is the prototype program. Evaluation of this stage should compare the prototype to other existing programs. Professionals involved in parent training will readily recognize the benefits and problems presented by this three-stage delineation. Parent demonstration training programs are frequently involved in all three stages simultaneously. Evaluation becomes a difficult organizational, administrative, and practical problem. In addition, the problems with ongoing operational programs, not previously evaluated, are identical in that they are in effect "demonstration" programs as well.

Stufflebeam (1971) noted the above problems as well as the controversy regarding the usefulness of experimental methodology in evaluating social change problems. Smith (1979) pointed out that for any social or educational program the question is not whether or not experimental design is a useful approach for evaluation, but whether experimental methodology is appropriately matched to the proper area of evaluation. Four types of decisions to be made by program developers and, subsequently, four types of evaluation were identified by Stufflebeam. Context evaluation is needed to identify and assess needs as well as the underlying problems. Input evaluation should assess program capabilities, input strategies, and designs for implementation. Process evaluation should record and monitor program activities. Product evaluation should determine if project objectives are met and attempt to determine the cause of evaluation

results (Smith, 1979). Stufflebeam concludes that experimental methodology has great relevance for product evaluations, less for input evaluations and no relevance for content or process evaluation.

We professionals involved in parent training need information for making all four types of decisions. We cannot afford to concentrate our energies on context and process evaluation to the detriment of input and product evaluation--or vice versa. We must attend to both. Traditionally, we have made evaluation decisions using non-experimental assessment procedures, and for many aspects of program evaluation we should continue to do so. However, we are still left with the problem of fitting our rectangularly-shaped parent programs into the round holes of experimental evaluation where appropriate.

The practical problems encountered in evaluating parent programs are numerous and frustrating. Random selection of parents as subjects is practically impossible since parents usually enter these programs as volunteers or members of a parent organization. Random assignment is also difficult due to number of participants and parent preference for particular programs. Parents present themselves with wide variations in knowledge and skills. What may be an in-depth program for one parent becomes simply a review or maintenance program for another, thereby, influencing pre- and posttest results. Program administrators are prone to making program changes before all the data have been gathered. Parent training is often conducted indirectly by mail or through the media and is difficult to evaluate. These are just a few of the problems plaguing experimental and non-experimental evaluations of parent training programs.

A number of procedures currently used could be employed in experimental, quasi-experimental, or non-experimental evaluation of parent programs. Kroth (1975) suggests using a daily report card system as an impetus for positive parent/child interaction and as a two-way communication vehicle between parent and teacher. This program could be implemented using a random selection of parents and a control group with pre and post academic measures taken on the children of both groups. Parent attitude scales such as Shoben's (1949) readily lend themselves to experimental pre/post evaluation designs and are appropriate for use with many types of parent programs. Attitude scales may also be used non-experimentally to screen participants or as follow-up measures to assess desired outcomes. Other measures

such as checklists, value clarification activities, and analysis of specific content learning may be utilized in experimental, quasi-experimental and non-experimental evaluation.

When preparing an evaluation plan, the evaluator should first consider why that particular parent program was selected for implementation. Does it complement a larger program such as a physically handicapped children's unit where, for example, parents are trained as physical therapists? Does it meet an identified need? Or is it one of many parent programs under evaluation?

Second, the evaluator should determine what the staff would like to be able to say about the parent program. Examples of this might be that the program was well received, that the children showed improvement in school performance, or that parental knowledge levels were substantially increased.

Third, the evaluator must formulate evaluation questions that will provide the necessary information. Questions might be: Did children of parents receiving training show improvement in reading tests? or Was the program effective in increasing parents' knowledge of behavior management techniques?

Fourth, it must be determined what evaluation procedure will most effectively provide the needed information. Here, one might stop to consider the potential recipients or audiences for the evaluation information. This may affect the overall evaluation design. Funding sources will probably be interested in the number of parents served, the number of handicapping conditions represented, the frequency of parent contact, the types of problems presented by parents, the changes in children's programs as a result of parental participation, and the types of programs presented or available to parents. Program administrators will probably be interested in the best time to involve parents, cost-effectiveness of programs, value of available resources (e.g., kits, books, and films), consumer satisfaction, characteristics of parents who are involved and, of course, child change information. Educators and parents are interested in parental attitudes, normative data on children (such as amount of play time, structured activity time, normal bed times, etc.), and the effectiveness of a variety of techniques to change child and parent behavior. An often neglected area of concern is the long-term effects of parent programs which can be

determined by follow-up questionnaires or observations. In addition, one may be interested in the generalization of training to other members of the family.

Finally, the functions that the evaluation results must be determined. Brooks (1965) suggests four potential functions: accountability, feedback, dissemination, and theory building. The evaluator should consider each step in the evaluation process carefully, before beginning to determine consistency and compatibility. Figure 1 is an example of a possible evaluation sequence.

While the difficulties involved in the evaluation of multifaceted parent programs may be tenacious, the goal of successful parent involvement necessitates continued evaluation efforts. These efforts will facilitate informed decision making by program evaluators concerned with program improvement and justification. The growth in parent involvement in recent years is encouraging, and it is the responsibility of the concerned professional to provide quality training to insure positive benefits to parent, child, and educator.

FIGURE 1
Evaluation Sequence

Program Choice: Training parents as behavior modifiers (requested by local parent group)

Concern	Evaluation Question	Method	Audience
1. Is this program effective for training parents?	Does parents' knowledge of behavior management techniques increase as a result of training?	Content test[a] specific to program, administered pre and post. Graphing[a] by teachers of children's behavior from baseline to conclusion of program, using a control group.	Funding agency Program staff

2.	Are parent attitudes affected by training?	Do parents show changes in attitudes toward parenting following training?	Parent attitude scales[b] administered pre and post. Self-evaluation forms.[c] Q-Sorts[d] administered pre and post.	Program staff Educators Prospective consumers Funding agency

[a]Cooper and Edge, 1978; [b]Shoben, 1949; [c]Colletta, 1977; [d]Kroth, 1972

APPENDIX
Parent Involvement Assessment Scale (PIAS)

INSTRUCTIONS. Listed below are a number of areas where parents and teachers can become intimately involved in the education of children. Rate each item from 1 (low) to 5 (high) as to whether it exists and how effectively or efficiently it operates. At the end of the list are some suggestions of activities that might influence your rating.

METHOD. As the building principal, rank each item from your perspective. Select a number of parents at school or ask each parent on parent conference day to fill out the PIAS and leave it in a specified box. Ask all teachers to respond to the PIAS. Aggregate the data from each group and place it on a master sheet. Check to see where discrepancies exist.

	Principal Rating	Parent Rating	Teacher Rating
Placement			
Program planning			
Advisory group			
Parent program			
Parent volunteer			
-Instructional activities			
-Non-instructional activities			
Progress reporting (evaluation)			
Home visitation program			
Parent/teacher organization			
Open communication			
Parent comfort			

CONSIDERATIONS FOR RANKING EACH ITEM

PLACEMENT. A high rating would indicate that almost all parents come and they feel free to bring a friend. They enter into the discussion and actually talk a fair share of the time. On occasion they express to others (outside the school and filtered back) about how well they felt they were treated and how appropriate they felt the placement was. Parents do not hesitate to suggest changes in placement and are involved in minor change decisions as well as major changes.

PROGRAM PLANNING. A high rating would indicate much more than having parents sign an IEP form. At the offset of a planning conference, parents are asked what goals or objectives they have for the child, which are written into the IEP. They share with teachers techniques (i.e., types of reinforcers) that seem to motivate the child. They have activities to do at home to complement the school program and they feel free to suggest changes in their child's program.

PARENT ADVISORY GROUP. A high rating would indicate more than a group of parents that meets two or three times a year to ask questions or receive information from the school administrators. The Parent Advisory Group (PAG) would have clearly delineated guidelines based on School Board policy and state laws, and the group membership would be representative. Meetings would be held regularly and be well attended. The PAG would have activities to carry out that would be active rather than just reactive. They might be involved in developing questionnaires, conducting surveys, organizing other parents and providing input to school personnel on attitudes and opinions of parents in the area. They would keep minutes of the meetings to be available to share with others.

PARENT PROGRAMS. A high rating would indicate that there were a variety of parent education programs available to the patrons of the school. Parents might be involved in conducting as well as participating in the programs. The programs would be conducted at a variety of times of day and throughout the school year and would be of such a scope that a majority of the parents would attend at least one of the programs.

PARENT VOLUNTEERS. Parents can be a valuable resource to a school in both instructional and non-instructional activities. Parents have been taught to

run programs such as the Monterey, Distar, and Frostig. They have been used for speaking to classes on special interests and for on-the-job training of older students. Parents can grade papers, duplicate materials, and aggregate classroom performance data. They can conduct fund-raising activities and work on school benefaction projects. A high rating would indicate that parents are actively involved in school activities and that there is a training program for volunteers.

PROGRESS REPORTING. A high rating would indicate that most parents attend pupil progress reporting conferences that are regularly scheduled. Released time would be available for such contacts and provisions would be made for weekly or daily reporting for selected cases. Inservice training would be available for the staff and for the parents on techniques for sharing information.

HOME VISITATION PROGRAM. A number of school districts have found home visitation programs an effective strategy to establish a bond between school and home. A planned program would receive a high rating. If home visitations are only reserved for severe problem cases, then this would result in a lower rating.

PARENT/TEACHER ORGANIZATION. A high rating would be given to the situation where the formal PTA or PTO has succeeded in involving parents and teachers in active interchanges. These meetings might involve small group discussions on discipline, progress reporting, sex education, etc. They might also include large group programs such as open house and all school programs.

OPEN COMMUNICATION INDEX. A high rating would indicate that there is a free flow of communication between school and home, and home and school. School personnel are probably involved in a planned "call a parent a week" program to report positive results. Personal notes, letters, and school news are passed to parents often. Parents feel free to call or visit the school at all times.

PARENT COMFORT INDEX. A high rating would indicate that parents feel comfortable in visiting the school. On any given day parents could be found in the lounge having a cup of coffee or visiting with the principal, secretary, counselor, librarian or teachers. While this category is somewhat subjective, the presence of parents on the school grounds is usually a good objective measure.

REFERENCES

Bernal, M. E., & North, J. A. A survey of parent training manuals. Journal of Applied Behavioral Analysis, 1978, 2(4), 533-544.

Brooks, M. D. The community action program as a setting for applied research. Journal of Social Issues, 1965, 21, 29-40.

Campbell, D. T. Considering the case against experimental evaluations of social innovations. Administrative Science Quarterly, 1970, 15, 110-113.

Coletta, A. J. Working together: A guide to parent involvement. Atlanta, GA: Humanics Limited, 1977.

Cooper, J. O., & Edge, D. Parenting: Strategies and educational methods. Columbus, OH: Charles Merrill, 1978.

Duncan, L. W., & Fitzgerald, P. W. Increasing the parent child communication through counselor-parent conferences. Personnel and Guidance Journal, 1969, 514-517.

Gordon, T. Parent effectiveness training. New York: Wyden, Inc., 1972.

Kroth, R. L. Communicating with parents of exceptional children: Improving parent-teacher relationships. Denver: Love Publishing Co., 1975.

Kroth, R. L., & Simpson, R. Parent conferences as a teaching strategy. Denver: Love Publishing Co., 1977.

Kroth, R. Target behavior. Olathe, KS: Select-Ed, 1972.

O'Dell, S. Training parents in behavior modification: A review. Psychological Bulletin, 1974, 81, 418-433.

Radin, N. The impact of a kindergarten home counseling program. Exceptional Children, 1969, 37, 251-255.

Schaefer, E. S. Parents as educators: Evidence from cross-sectional, longitudinal and intervention research. Young Children, 1972, 27, 227-239.

Shearer, M. S., & Shearer, D. E. The Portage Project: A model for early childhood education. Exceptional Children, 1972, 39, 210-217.

Shoben, E. J., Jr. The assessment of parental attitudes in relation to child adjustment. Genetic Psychology Monographs, 1949, 39, 101-140.

Smith, M. D. Experimental assessment of social programs. Catalog of Selected Documents in Psychology, 1979, 9(M. no. 1801).

Stufflebeam, D. L. The use of experimental design in education evaluation. Journal of Educational Measurement, 1971, 8, 267-274.

Suchman, E. A. Action for what? A critique of evaluative research. In C. H. Weiss (Ed.), Evaluating action programs. Boston: Allyn & Bacon, 1972.

Wagonseller, B. R., Burnett, M., Salzberg, B., & Burnett, J. The art of parenting. Champaign, IL: Research Press, 1977.

Weiss, R. S., & Rein, M. The evaluation of broad-aim programs: Experimental design, its difficulties, and an alternative. Administrative Science Quarterly, 1970, 15, 97-109.

Evaluating Staff Development

Linda L. Lynch and Valerie Lynch

As HCEEP projects are developed, major attention is focused on child services and family involvement. Concurrently, plans are developed for evaluation, demonstration, and dissemination. With the exception of providing some inservice training to staff, planning for a staff development program is rarely addressed. Why does staff development receive so little attention?

At one point or another, most administrators in the field of special education have been direct service providers--educators, speech-language pathologists, occupational therapists, and so on. Service providers focus on children and families as clients and, further, measure success through documentation of child progress.

When administrative responsibilities are assumed, the emphasis on child progress continues, even though administrators have little direct impact on children. But the client population has changed--and has become the staff. Only by focusing on staff development can administrators hope to impact on child learning.

From a direct service provider's standpoint, one can also experience problems in the area of staff development. How much information on staff development is received during the college training years? How many people are taught the importance of and ways of influencing staff development programs? It seems that little, if any, information is provided in this area. That is, professionals generally have never been taught to rely on staff development activities as critical elements in providing better (or the best) services to children and their families.

It is up to program administrators to take a new look at staff development. The staff development component may be the major "product" of HCEEP projects, an important reason for its re-evaluation.

RATIONALE (OR, WHY BOTHER WITH STAFF DEVELOPMENT?)

Before describing a model for staff development, it seems appropriate to delineate reasons for initiating a staff development program. The following objectives were generated in discussions with a number of administrators and other professionals:

1. To improve staff skills and performance.
2. To improve services to children and their families.
3. To provide the basis for dissemination of specific program components.
4. To improve communication among staff.
5. To provide reinforcement and improve morale.

A MODEL FOR STAFF DEVELOPMENT

Staff development is an arranged system for improving and/or changing attitudes, knowledge, and skills of staff members in order to upgrade their ability to perform specific tasks. What is that "arranged system"? A model of staff development is presented in Figure 1 which may be individualized to a variety of situations and which can provide the structure in which staff development occurs, i.e., the arranged system. Each of the steps in the process are described.

FIGURE 1
Staff Development Model

1. Define Role Responsibilities (Product)
2. Establish Competencies for Each Role Responsibility (Processes)
3. Assess Needs
   - 3.1 Related to competencies
   - 3.2 Emergent needs
4. Develop Goals & Objectives
   - 4.1 Attitudinal/affective
   - 4.2 Content/knowledge/cognitive
   - 4.3 Performance/motoric
5. Develop Activities
   - 5.1 What
   - 5.2 Who is responsible
   - 5.3 By when
6. Develop Evaluation Procedures
   - 6.1 Attitudinal/affective
     - 6.1.1 Product
     - 6.1.2 Process
   - 6.2 Content/knowledge/cognitive
     - 6.2.1 Product
     - 6.2.2 Process
   - 6.3 Performance/motoric
     - 6.3.1 Product
     - 6.3.2 Process
7. Implement Activities
8. Evaluate
   - 8.1 Process
   - 8.2 Product

1. DEFINE ROLE RESPONSIBILITIES. Most administrators define role responsibilities but, unfortunately, it is often done very loosely and after an individual has been hired to fill the role. In order to meaningfully define role responsibilities, it is important to define the "products" of each role or staff position. For example, a product of a teacher might be defined as "child learning" or "curriculum development." A product of an

administrator might be "staff learning" or "an evaluation system." A product of clerical staff might be "systematic filing procedures" or "functional paper flow." Until products are defined with some specificity, it is impossible to define the processes by which the products will be developed.

2. ESTABLISH THE PROCESSES OR COMPETENCIES REQUIRED FOR EACH ROLE RESPONSIBILITY. This step involves defining what a given individual needs to know in order to produce what is expected (i.e., the product). Although this is a logical step, all too often it is relegated to a level of generalities or is confused with qualifications (for example, "master's degree and two years experience" or "typing speed of 90 words per minute" or "ability to relate to people"). Do those "competencies" mean that an individual can be a teacher or a secretary or an administrator? How do those qualifications relate to the product expected?

The Wisconsin Department of Public Instruction (1977) has developed a Teacher Performance Evaluation Instrument which defines the competencies felt to be necessary for performance as an Early Childhood-Exceptional Education Needs (EC-EEN) teacher. For each of ten products (e.g., assessment, curricular strategies, instructional activities), a number of critical processes are defined. An excerpt of the competencies defined for Instructional Activities is included in Figure 2. In this example, as in others, it is important to note that competencies are individualized to specific educational settings. That is, competencies must be individualized to a particular role in a particular school or project. What works well in Wisconsin may not fit the needs of a project in California.

Specifying competencies for roles is often facilitated by involving the persons performing in those positions or similar positions. Involvement in this step can assist staff in the identification of their individual and group strengths and weaknesses as well as the project's needs for certain performance levels. This can enhance morale and relieve the anxiety often associated with evaluation of competencies.

FIGURE 2
Teacher Competencies

Competency #2 - Instructional Activities

1 = I have no knowledge or skill.
4 = I have an awareness of or skill.
7 = I have working knowledge (i.e., understanding) and/or ability to use with assistance and/or additional information.
10 = I have expert knowledge and/or skill to implement with proficiency.

Early Childhood-Exceptional Educational Needs teachers... will be able to:

Utilize applied behavior analysis technology (e.g. reinforcement, recording, task analysis, etc.).	1...4...7...10
Organize classroom space, equipment, and materials to accomodate individual objectives.	1...4...7...10
Modify instructional activities for specific handicapping conditions.	1...4...7...10
Modify instructional activities to individual learning styles of children.	1...4...7...10
Record and report child's program and progress for allied professionals, paraprofessionals, and parents.	1...4...7...10

Note. Excerpt from Early Childhood-Exceptional Education Needs (EC-EEN) Teacher Performance Evaluation Instrument (Rev. ed.). Wisconsin Department of Public Instruction (SIG), 1977.

3. ASSESS NEEDS. Once competencies have been defined, staff needs must be assessed. The needs assessment takes two forms: (1) the assessment of each staff person's needs relative to the competencies established for that position; and, (2) the assessment of emergent needs, i.e., needs not specifically related to the competencies but affecting the outcomes of the position.

A needs assessment not only identifies staff needs, but can also minimize the discrepancy between the staff vs. supervisors or trainees vs. trainers perceptions of what ought to be taught or supervised. The needs assessment serves two purposes. First, it allows a supervisor to focus on his or her perception of needs as well as the experiences and interests of those who will be involved in the staff development program. Second, a needs assessment encourages staff members to analyze their present skills, needs, and the resources available to them in their own situation.

There are several methods for assessing the needs of staff--through a written instrument, an individual interview, a group meeting, or a combination of these methods. The Wisconsin Teacher Performance Evaluation Instrument (described above) could also be used to assess staff needs. With such an instrument, it would be possible to have each individual do a self-needs assessment. And, the instrument could be used by a supervisor to observe areas in which he or she perceives needs. Interviews or group discussions around the competencies delineated in the instrument can provide pertinent information regarding strengths and needs. Another tool for assessing staff needs for inservice training sessions is included in the Appendix. It can be periodically administered to assess the changing needs of staff and modified to accurately reflect the major areas of staff concern.

The important point about the needs assessment is that one should differentiate between and acknowledge (1) needs related to competencies and (2) emergent needs. Further, each staff person must have the opportunity for input, for without input the staff development program will become an imposition instead of a meaningful process.

4. DEVELOP GOALS AND OBJECTIVES. Once needs have been assessed and ranked by priorities, it is time to develop goals and objectives for staff development. Goals and objectives may be in one or all of these domains: (1) attitudinal/affective, (2) content/knowledge/

cognitive, and/ or (3) performance/motoric. Generally, the goals will identify and describe general areas of training. The objectives will identify and describe specifically what the participant will have learned at the end of a training event. See Figure 3 for an example of the specification of goals and objectives.

The process of developing goals and objectives necessitates a discussion or negotiation between the individual staff person or staff as a group and the supervisor/administrator. The supervisor must be open to individual needs and be creative in assisting the staff to develop goals and objectives which are reasonable. The staff must also be open, creative, and reasonable, realizing that resources are limited. Critical to this step is consideration of criteria by which one can determine whether the goals and objectives have been met. These criteria, in fact, will define some of the procedures to be employed in evaluating staff development.

5. DEVELOP ACTIVITIES. The development of activities to accomplish the stated goals and objectives is also a part of the negotiation process described in the previous step. The outcomes of that negotiation process can be displayed in writing as an agreement between the staff person and supervisor. A sample staff development agreement is presented in Figure 4. This agreement serves as the basis for activities throughout the staff development program. The agreement defines the parameters of the staff development program by articulating the following:

a. goals
b. objectives
c. activities
d. evaluation procedures
e. timelines (by when)
f. responsibilities (who is responsible)

FIGURE 3
Establishing Staff Development Goals and Objectives

Goal:	Identify and describe general areas of training.
Specific Objective	Identify and describe specifically what the participants will have learned at the end of a training event.
Criteria:	1. Observable and Measurable. In order to determine whether or not the desired learning has taken place, participants may be asked to do something which demonstrates what was learned. Specific objectives should be phrased in terms of active verbs. For example: to list, to write, to define, to demonstrate. 2. Specify a Level of Acceptable Performance. To ensure the quality of the participant's performance following staff development activities, the objectives must indicate in quantitative terms what the participant must do to demonstrate what has been learned. The importance of this criterion is that it helps the staff development person focus on the design and selection of training methods. 3. List Important Conditions for Performance. The conditions and restrictions under which the participants will perform are noted to assist in the selection of techniques and methods to accomplish the training and promote learning. For example: conditions may be established which require working with a partner or using handouts as study guides.

Adapted from: Hickey, J., 1976.

FIGURE 4
Staff Development Agreement

Goal _____

Objective _____

Special Activities:	By When (Date)	Who is Responsible
_____	_____	_____
_____	_____	_____

Evaluation Method:	By When (Date)	Who is Responsible
_____	_____	_____
_____	_____	_____

Staff Person _____ Date _____

Supervisor _____ Date _____

It is important to note that inclusion of timelines and persons responsible will increase the likelihood of the program being carried out. It is also notable that use of this strategy--development of the agreement itself--further minimizes discrepancies between staff and supervisors' perceptions of what ought to be taught or supervised in that it requires continued communicative interaction from all participants.

6. DEVELOP EVALUATION PROCEDURES. For emphasis, developing goals and objectives, activities, and evaluation procedures are separate steps in the model. The development of the evaluation procedures should, however, also be a part of the negotiation process described in the development of goals, objectives, and activities (Steps 4 and 5 above) for staff development.

Evaluation is the process of determining what kind of an impact staff development activities have on the behavior of the individual. That is, how does one determine whether there was an impact and, further, the

magnitude of that impact? All too often, evaluation of staff development activities consists of counting the number of people and the number of inservice training activities in which they participate. These kinds of data may be of some use but they do not provide any information regarding resulting behavior changes. It is absolutely essential to determine the strategies for evaluation prior to the implementation of any kind of staff development activity. In so doing, one can assure the gathering of appropriate information which will accurately detail the impact of the staff development activities. If careful attention has been given to the development of goals and objectives which: (1) are observable and measurable, (2) specify a level of acceptable performance, and (3) list important conditions for performance, then the basis for evaluation methods or procedures will already have been articulated.

Consider the example of evaluation procedures in Figure 5. In this example, three objectives specify the basis of the evaluation procedures. Objective 1 states that "teachers will be able to list..." How does one evaluate whether or not this objective was met? Ask the teachers "to list" and "grade" (evaluate) applicable measurements. That information should determine: (1) which teachers met the objective and (2) which teachers did not meet the objective. Critical at this point are: (1) what steps will now be undertaken--what new (additional) objectives can be developed to assist teachers who did not meet the criterion (i.e., did not list four types of measurement); and (2) what steps will be taken to reinforce or provide positive feedback to those who did meet the criterion.

Objective 2--"teachers will be able to write..." --delineates what is necessary to the accomplishment of the objective. As with the first objective, someone needs to "grade" or evaluate the written plans. Part of establishing evaluation procedures is to determine who is responsible and how that will be done. In this instance, a trainer or supervisor may assign points to each aspect requested in the objective, or perhaps the teachers could review the plans and determine whether or not the criterion was met. The information gathered should indicate (1) which teachers met the objective, and (2) which teachers did not meet the objective. Again, that information should be used to assist teachers who did not meet the criterion and to reinforce those who did.

FIGURE 5
Including Evaluation Procedures
in Staff Development Objectives

Assess and Identify Needs	Teachers have difficulty in using on-going measurement. Need to learn about measurement systems and their application to instructional programs.
Develop Goals and Objectives	
Goal:	Teachers will understand the various types of measurement and be able to incorporate them into instructional plans.
Objective 1:	At the end of a 2-hour lecture, teachers will be able to list, without the use of notes, four types of measurement applicable to instructional plans.
Objective 2:	At the end of a 2-week period, teachers will be able to write instructional plans which include a specified type of measurement system for at least two children in their classroom.
Objective 3:	At the end of a 3-week period, teachers will be able to implement an instructional program which includes a specified type of measurement system for at least two children in the classroom.

Objective 3 details an implementation activity of a specific nature--"includes a specified type of measurement system." Again, how does one know if the objective has been met? Perhaps an observation form could be developed so that the supervisor could make observations in the classroom to determine whether the criterion had been met. Perhaps the measurement information collected by the teacher could be submitted to the supervisor for review. The point is that the basis for evaluation is present in the objective. As an example, the Teaching

Research Infant and Child Center, Monmouth, Oregon, has an observation form which details a number of behaviors that can be observed and upon which feedback can be given.

In addition to those evaluation methods specified in the staff development agreements, the following questions can be helpful in shaping staff development evaluation methods. There may be a variety of ways to answer the questions, but their consideration prior to the implementation of activities can help determine the evaluation procedures.

1. To what extent were staff development objectives accomplished?
2. What is the short-term and the long-term impact of the staff development program?
3. How well was the staff development program conducted?
4. What was done well and what was done poorly?
5. What changes could or should be made to improve the staff development program?
6. Did participants acquire new skills and are they able to use those skills on the job?

7. IMPLEMENT THE ACTIVITIES DELINEATED IN THE STAFF DEVELOPMENT AGREEMENT. The critical issue here is the monitoring of activities by those who are listed as being responsible. Staff development programs often become "lost" or ineffective because no one has responsibility for ensuring the implementation of activities.

8. EVALUATE. Evaluation is an ongoing process. As stated by Bishop (1976), evaluation should contribute to decision making and in-process corrections to program improvement, to reporting and feedback, and to creativity and variety in the inservice efforts. Evaluation is not research, testing, measurement, or statistics, although they will probably be included. Evaluation should provide information on the efficiency of the staff development program as a whole. However, it should also provide information which allows modification of each step in the model in order to create a better staff development program. In other words, evaluation procedures should provide information on both the product and the process.

It seems pertinent at this time to point out how this staff development model might also form the basis of an HCEEP demonstration project's demonstration and dissemination activities. If a given program wishes to demonstrate and disseminate the services to children component, important aspects are the personnel and their delivery of services within that component. If the staff development activities of the project have followed a model such as that presented, training procedures can be developed and implemented around specific personnel roles and the competencies deemed necessary to deliver the products or outcomes--services to children. The staff development model and the accompanying training procedures become a "package" to be used in demonstration and dissemination activities.

SUMMARY

Staff development is critical to successful program development and implementation. Re-evaluation of this component in light of the products desired in a given program may, in fact, have a significant impact on the success of the program. As one of the participants at this workshop commented, "The model looks like an IEP for staff." In fact, that is what it is. This staff development model requires the same careful consideration, staff commitment, and staff involvement as the development of IEPs for children. This consideration, commitment, and staff involvement can only result in better (or the best) services to handicapped children and their families.

APPENDIX
Needs Assessment for Inservice Training

GENERAL AREAS

Directions: Please check the three areas which you feel are your highest priorities for training in the area of the severely handicapped.

_____ 1. Initial assessment procedures for pupils/clients
_____ 2. Establishing instructional goals and short term objectives
_____ 3. Planning pupil/client programs
_____ 4. Selection of program materials
_____ 5. Selection of program strategies
_____ 6. Use of ongoing measurement procedures
_____ 7. Overall evaluations of pupil/client progress (e.g., quarterly, annual)
_____ 8. How to work cooperatively with other educational professionals
_____ 9. Management of professional time
_____10. Inservice presentations from other professionals (e.g., OT, PT, Speech Therapist, Social Worker, Nurse, Psychologist, Educator)
_____11. Identifying and locating resources
_____12. Training others (paraprofessionals, parents, professionals)
_____13. Scheduling, coordinating and managing resources
_____14. Designing, managing and administering special programs (e.g., infant, vocational, parent training)
_____15. Legal rights of the handicapped
_____16. Other:
_____17. Other:

SPECIFIC AREAS

Directions: Please check the two areas which you feel are your highest priority areas for training under each of the following categories. Check only two in each category.

1. INITIAL ASSESSMENT PROCEDURES FOR PUPILS/ CLIENTS

_____1.1 Assessment instruments
_____1.2 Assessment techniques
_____1.3 Assessment rationale
_____1.4 Other:

2. ESTABLISHING INSTRUCTIONAL GOALS AND SHORT-TERM OBJECTIVES

_____2.1 Appropriateness
_____2.2 Rationale
_____2.3 Format (how to)
_____2.4 Priority ranking of goals and objectives
_____2.5 Other:

3. PLANNING PUPIL/CLIENT PROGRAMS

_____3.1 Pinpointing behaviors
_____3.2 Task analysis
_____3.3 Program format (how to)
_____3.4 Other:

4. SELECTION OF PROGRAM MATERIALS

_____4.1 Identifying available materials
_____4.2 Evaluating materials
_____4.3 Modifying existing materials
_____4.4 Other:

5. SELECTION OF PROGRAM STRATEGIES

_____5.1 Steps in the learning process
_____5.2 Determining what strategies are appropriate
_____5.3 Behavior management
_____5.4 Other:

6. USE OF ONGOING MEASUREMENT PROCEDURES

_____6.1 Selection of tools and procedures
_____6.2 Implementing measurement systems
_____6.3 Use of measurement information (i.e., to modify programs)
_____6.4 Other:

7. OVERALL EVALUATIONS OF PUPIL/CLIENT PROGRESS (e.g., QUARTERLY, ANNUAL)

_____7.1 Selection of tools and procedures
_____7.2 Implementing evaluation (how often)
_____7.3 Use of the evaluation information (e.g., future programming, grouping pupils)
_____7.4 Other:

8. HOW TO WORK COOPERATIVELY WITH OTHER EDUCATIONAL PROFESSIONALS

_____8.1 Communication techniques
_____8.2 Roles and descriptions
_____8.3 Implementing a team approach to providing services to pupils
_____8.4 Other:

9. MANAGEMENT OF PROFESSIONAL TIME

_____9.1 Ranking activities by priority
_____9.2 Scheduling priorities
_____9.3 Coordinating schedules with other professionals
_____9.4 Scheduling classroom/client programs
_____9.5 Other:

10. INSERVICE PRESENTATIONS FROM OTHER PROFESSIONALS
(List professionals by numbers from the list below.)

_____10.1 Assessment_____
_____10.2 Program planning_____
_____10.3 Measurement, evaluation_____
_____10.4 Instructional strategy_____

_____10.5 Resources _____

1. Teachers/Educators
2. Administrators
3. Speech Clinicians
4. Physical Therapists
5. Occupational Therapists
6. Family Services/ Social Workers
7. Nurses
8. Vocational Educators
9. Psychologists
10. Other:

11. IDENTIFYING AND LOCATING RESOURCES

_____11.1 Identifying the need for resources
_____11.2 Identifying community resources
_____11.3 Selecting appropriate resources
_____11.4 Other:

12. TRAINING OTHERS (PARAPROFESSIONALS, PARENTS, PROFESSIONALS)

_____12.1 Parents and families
_____12.2 Volunteers
_____12.3 Assistant teachers, OT's, etc.
_____12.4 Organize an inservice (how to)
_____12.5 Other:

13. SCHEDULING, COORDINATING AND MANAGING RESOURCES

_____13.1 Use of resource personnel in the classroom
_____13.2 Working in and with the community
_____13.3 Public relations
_____13.4 Parent groups
_____13.5 Other:

14. DESIGNING, MANAGING AND ADMINISTERING SPECIAL PROGRAMS

_____14.1 Infant programs
_____14.2 Vocational programs
_____14.3 Community living programs
_____14.4 Family training
_____14.5 Other:

15. LEGAL RIGHTS OF THE HANDICAPPED

_____15.1 P.L. 94-142
_____15.2 504
_____15.3 State rules and regulations
_____15.4 Other:

16. OTHER: (Refer to the general needs assessment and specify)

16.1 _____

16.2 _____

17. OTHER: (Refer to the general needs assessment and specify)

17.1 _____

17.2 _____

If you have any questions or concerns, please contact us:

Center for Inservice Training and
Program Development
University of Washington, JD-05
Seattle, Washington 98195

REFERENCES

Bishop, L. J. Staff development and instructional improvement. Boston: Allyn and Bacon, 1976.

Hickey, J. A guide to inservice training. Unpublished manuscript, Center for the Severely Handicapped, University of Washington, 1976.

Teaching Research Infant and Child Center. Training in the Teaching Research data based classroom model. Monmouth, OR: author, undated.

Wisconsin Department of Public Instruction (SIG). Early childhood exceptional education needs (EC-EEN) teacher performance evaluation instrument (Rev. ed.). Madison: Author, 1977.

ADDITIONAL READINGS

Bean, R., & Clemes, H. Elementary principal's handbook: New approaches to administrative action. West Nyack, NY: Parker Publishing Company, 1978.

Bolton, D. L. Selection and evaluation of teachers. Berkeley, CA: McCutchan Publishing Corporation, 1973.

Dunn, R., & Dunn, K. J. Administrator's guide to new programs for faculty management and evaluation. West Nyack, NY: Parker Publishing Company, 1977.

English, F. W. School organization and management. Worthington, OH: Charles A. Jones, 1975.

Hack, W. G., et al. Educational administration: Selected readings. Boston: Allyn and Bacon, 1971.

Hersey, P., & Blanchard, K. H. Management of organizational behavior: Utilizing human resources (2nd ed.). Englewood Cliffs, NJ: Prentice Hall, 1972.

Knudson, H. R. Human elements of administration. New York: Holt, Rinehart and Winston, 1963.

Lynch, V. W. It's the principal of it all: An overview of the development of a personnel system. Unpublished manuscript, University of Washington, 1979.

Melcher, J. W., & Franks, D. J. Certification and training of teachers of young handicapped children: The Wisconsin experience (Bureau Memorandum, Vol. 20, No. 1). Madison: Wisconsin Department of Public Instruction, no date.

Moeller, G. H., & Mahan, D. J. The faculty team: School organization for results. Chicago: Science Research Associates, 1971.

Sergiovanni, T. J., & Carver, F. D. The new school executive: A theory of administration. New York: Harper & Row, 1973.

Townsend, R. Up the organization: How to stop the corporation from stifling people and strangling profits. New York: Fawcett World Library, 1970.

Evaluating the Impact of Demonstration and Dissemination Activities

Paul A. Bradley

This paper is concerned with evaluating the impact or effect of the demonstration/dissemination activities of an HCEEP project. The first part of the paper is concerned with the nature and critical components of impact evaluation studies. The second part of the paper is concerned with conditions that should be met before conducting an impact evaluation study of demonstration and dissemination.

NATURE OF IMPACT EVALUATION

An impact evaluation study is one whose basic purpose is to determine if the program is responsible for producing changes or causing effects. The purpose of an impact evaluation is quite unique and specific, unlike other types of evaluation studies with various purposes (e.g., formative, summative, or process evaluation). The unique purpose of impact evaluation makes it closely identifiable with experimental research, and the methodology of impact evaluation studies is similar to that of experimental research studies. While some evaluation specialists have argued against imposing experimental research methodology on evaluation studies, it has to be recognized that there are certain occasions when the purpose of the evaluation study requires the methodology of experimental research. With regard to the demonstration and dissemination (D&D) component of an HCEEP project, an impact evaluation study would try to determine if D&D activities are responsible for producing changes.

Let me try to indicate how an impact evaluation study of the D&D component stands apart from other evaluation studies of the D&D component. I consider the D&D component as a program by itself (as well as being a part of the total HCEEP project) with its own antecedents, transactions and outcomes, where the meaning of these terms is that originally given by Stake (1967). The transactions of the program are those communication activities which comprise the substance of the D&D component. A good deal of evaluation activity can go into judging the quality of these transactions. For instance, if one of the demonstration activities is to have a training workshop, then it would be quite reasonable to do an evaluation study of the workshop, and this evaluation study might focus on these questions:

-Did the participants like the workshop?
-Were the leaders of the workshop well prepared and knowledgeable?
-Did the educational activities of the workshop follow sound educational principles and practices?
-Did the workshop address the needs of the participants?

An evaluation study that attempts to answer these questions might be called a formative evaluation. But, in an impact evaluation, your primary question would focus on the outcomes of the program:

-Did the communication activity produce a change in the audience that received the communication?

In the example of the training workshop evaluation, participants may be given a test and/or an attitude scale at the end of the workshop concerning the education of young handicapped children. Is this an impact evaluation study? The answer is either no or a qualified yes. The critical factor is whether, or how well, you can say the workshop produced a change in the participants' level of knowledge about or attitude toward young handicapped children when they were back on the job. So, while this evaluation activity is concerned with outcome, it doesn't do a very good job of determining if the workshop produced a change in the participants' knowledge or attitude. An example of a better impact evaluation of this

workshop would be to measure participants on-the-job behavior before and after the workshop, noting any attitudinal or interactional changes that have taken place in the participants.

CRITICAL COMPONENTS OF IMPACT EVALUATION

If you plan an impact evaluation study of D&D activities, then a very important concern is how you do it. For several reasons, I am not going to take a cookbook approach toward conducting an impact evaluation study. A major reason is that you can't prescribe one set of procedures for an impact evaluation study that would work for all D&D programs. The impact evaluation will have to be tailored to the particular D&D program. A second major reason is that I want to focus on the conditions that should be met before you do an impact evaluation study. A third major reason is that it would take a great deal of time to say all there is to say about impact evaluation. For further discussion on impact evaluation, see Weiss (1972).

What I do want to point out are critical technical aspects of impact evaluation studies. These four aspects are research design, measurement, sampling, and data analysis.

RESEARCH DESIGN. While all four aspects of impact evaluation are critical, the research design of the study is the most important. The purpose of an impact evaluation study is to determine if the D&D communication activities are responsible for producing changes in the audience that received the communication. One element of this purpose is detecting a change in the audience, which requires you to measure the audience before and after the communication, or to measure the audience after the communication and compare it to another that didn't receive the communication. However, detecting a change in the audience and some other group is not all there is to the research design. You also need to be able to say with confidence that the D&D activities, rather than some other agent or event, are responsible for producing the change. Whether a research design permits you to conclude that the D&D activities have produced the change depends on the internal validity of the design. Campbell

and Stanley (1966) have written not only about the concept of internal validity but also about the variety of designs--experimental and quasi-experimental--that are possible. This reference is a valuable resource for the research design options for an impact evaluation study. A recent book by Cook and Campbell (1979) should also be an important resource. In planning your study, it probably would be wise to consult with a research methodologist who is knowledgeable not only about experimental design but also about quasi-experiments. If you are familiar with these two references, then you can communicate clearly with your evaluation consultant.

MEASUREMENT. Since impact evaluations are empirical, data collection is involved and that means some form of measurement. There are two important parts of the measurement process. One is the instrument you will use to measure the outcome or dependent variable(s). The other part is concerned with the reliability and validity of the data that you get from these instruments. The instruments you use to collect your data could take several forms, depending on the outcome variables you want to measure. The instrument might be a test of knowledge about the problems and potentials of young handicapped children, a questionnaire for measuring attitudes toward young handicapped children, an interview to assess recognition and awareness of your program, or an instrument as simple as counting the number of requests (written and phone) for information about your program.

A decision that has to be made concerning instruments is whether to use existing instruments or to develop your own. Since you may need different types of instruments, you may want to talk to several measurement persons. An educational measurement person could help with a test of knowledge about young handicapped children, but you may need to talk to a marketing research person to get help in developing the interview to assess recognition and awareness of your program. The technical issues of reliability and validity must also be considered. Kerlinger (1973) is a good starting point to become familiar with the various aspects of measurement.

SAMPLING. Since D&D could involve communication activities directed at audiences in the community, state and beyond, it is quite possible that your impact evaluation study could involve measurements of a sample of persons from those audiences. It is very important

that the persons you measure are representative of the total audience. Achieving a representative sample involves some form of a random sample. While the concept of a random sample is fairly simple, designing and conducting a random sampling plan can get complicated. Sociological researchers, market researchers, or a polling organization may be able to provide valuable assistance for sampling on a community or state level.

DATA ANALYSIS. Data analysis is what you do with the numbers you obtain through the measurement process. (This is also known as statistics.) The kinds of data analyses you do are a function of the research design and the number of outcome variables. If the research design is a true experiment and there is only one outcome variable, then the data analysis is quite simple and straightforward--some form of an analysis of variance. But, if the research design is some form of time series, then the data analysis gets more complicated (Glass, Willson, & Gottman, 1975). If there are several outcome variables, then the data analysis becomes multivariate (Tatsuoka, 1971). In the latter two cases, you may need the advice of a versatile data analyst.

CONDITIONS FOR IMPACT EVALUATION

This section will focus on some considerations to be made before deciding to conduct an impact evaluation.

1. Full-scale impact evaluation studies can get costly in terms of money and time.
2. Certain conditions usually need to be met in order for an impact evaluation study to show that the program is responsible for producing any changes. If these conditions are not met, chances are slight that an impact evaluation will show that the D&D activities are responsible for producing changes in the audience. Conditions to be met include:
   - 2.1 The program being evaluated is relatively stable.
   - 2.2 The program being evaluated has clearly defined goals and clearly defined processes that should lead to the accomplishment of goals.
   - 2.3 Valid measures of the outcome variables are available.
   - 2.4 An internally valid research design is possible.

Each of these conditions is important, and failure to meet any one of them could undermine the impact evaluation study (in the sense that the results of the study show that the D&D activities did not produce any changes).

STABILITY

The importance of condition 2.1 is that a program needs a certain amount of time to be operating before it is reasonable to expect that it is producing the desired changes. If the D&D activities are still in a formative stage of development, these unfinished products will not stand the test of an impact evaluation. An impact evaluation is a stern test of a program's effectiveness, so be fair to your D&D activities and let them evolve to their finished form (using the results of other evaluation studies to aid the development process) before subjecting them to this type of test.

DEFINED GOALS AND PROCESSES

My arguments for the importance of condition 2.2 are based on logic, reason, and persuasion rather than empirical support. If you want D&D activities to produce results for you, then you have to know where you're going and how you're going to get there. This is what a theory or model of D&D can help you achieve. There are two important elements to this condition. One element is to know what you want to achieve. An impact evaluation study shows that a program has produced changes by collecting data on outcome variables, so it is necessary to know what outcome variables to measure. If the D&D activities do not have a specific goal(s), then it may be very difficult to identify the outcome variables that would show the effectiveness of the activities. For instance, it seems that it is quite possible to develop a D&D component composed of a variety of well designed and conducted activities that demonstrate the program to on-site visitors and disseminate information about the program and young handicapped children. But, carrying out an impact evaluation that stands a good chance of showing that the D&D activities produced changes would be rather difficult, because it is not known what outcome

variables should be measured. Choosing the outcome variables for this study is like Russian roulette--several outcome variables could be used, but maybe only one or two are variables on which the program has had an impact. A possible solution is to measure all relevant outcome variables, but there are problems in doing that. One problem is having valid measures of all the relevant outcome variables; another is cost, in terms of time and money. In impact evaluation studies, there usually are only a small number of outcome variables. So, the problem (possibly a dilemma) when there aren't specific goals for D&D activities is one of deciding which outcome variables, from all relevant variables, will be selected for measurement. This problem is avoided or minimized when the D&D activities are planned with specific goals (or outcomes) in mind.

The second important element to condition 2.2 is that, given specific goals (or outcomes), the D&D activities are chosen because they are capable of producing the outcomes. The capability might be known through experience (empirical data) or suggested by theory. In any case, the D&D activities need to be coordinated with the desired outcomes--every D&D activity should contribute in some known or suspected way to the achievement of goals. Very definitely this is a means-ends view that places the ends on a higher priority than the means. D&D activities are chosen because of their instrumentality in achieving the goals--not because of their elegance, not because of their simplicity, not because they are activities that you know how to do. While the qualities of elegance, simplicity, and expertise are important reasons for choosing a particular D&D activity, they are secondary to their value in producing the desired outcomes. Thus, if your goal is to increase the quality and quantity of services for young handicapped children in your community and state, you can conduct a task analysis of how that change can occur and plan the D&D activities that will accomplish the intermediate steps identified by the task analysis. That is, you need a "model" of how to use dissemination to increase the quality and quantity of services for young handicapped children.

The consequences of failing to have D&D activities clearly articulated with goals is that, if the activities are not chosen because of their capability of producing the desired outcomes, then they probably won't produce those

outcomes. If the activities don't produce the outcomes, an impact evaluation study will show that there are no changes on the outcome variables--no matter how technically sound the study is.

The moral of condition 2.2 is something like Snoopy looking for the Red Baron: Snoopy searches the northern sky but the Red Baron comes in from the south and Snoopy is shot down. If you want to find impact, you've got to look in the right place. Looking in the right place is significantly enhanced if the D&D program is conceived with specific outcomes in mind and the D&D activities are planned accordingly. By doing so, you can minimize the chances of your impact evaluation showing NSD (no significant difference).

VALID OUTCOME MEASURES AND RESEARCH DESIGN

Conditions 2.3 and 2.4 refer to technical qualities of the impact evaluation study per se, rather than to the D&D component. While the issues of measurement validity and research design are always important in evaluation studies, they are of critical importance in impact evaluation studies. Failure to achieve measurement validity and an internally valid research design will undermine an impact evaluation study. You could have a D&D program that actually has produced changes on the outcome variables, but if your instruments don't measure the variables, then you have the undesirable situation of having bad data--you wanted to measure apples but instead you measured oranges. The consequence, of course, is that there might not be any changes in oranges, and you reach the erroneous conclusion that your D&D has not had any impact. Frustrating? You bet!

Equally frustrating is the situation where you have an effective D&D program and valid measures of the outcome variables, but you have a research design that is weak in internal validity. For instance, you might have some valid measures of the quality and quantity of services for young handicapped children in the state taken before and after your D&D activities. Let's also say that these data show an increase in services from pre to post--a definite change in the outcome variables. But, alas, the research design is relatively weak and there are

reasonable challenges to the conclusion that the D&D program is responsible for producing the changes. In your heart and soul you know that the D&D activities produced the changes, and the data are right there for everyone to see, but you can't rule out competing hypotheses that something else besides your D&D program was responsible for the changes. Maybe it was your program, maybe it was a change that was going to happen anyway (regardless of your program) or maybe the change was produced by a combination of factors. The usefulness of good research design is that it eliminates many or all of the plausible competing hypotheses, leaving relatively unchallenged the hypothesis that the D&D component produced the change.

SUMMARY

This paper has addressed impact evaluation of the demonstration and dissemination component. Rather than focusing on the how-to-do-it aspects of impact evaluation, I have chosen to deal with the nature and critical components of impact evaluation and the conditions that should exist before you decide to do an impact evaluation.

When conducting an impact evaluation study, there are certain technical aspects of the study (design, sampling, measurement, data analysis) that require careful and expert attention. Positive results from an impact evaluation study are very useful in convincing skeptics of the program (as well as yourself) that the D&D activities have made a difference.

If you choose that it is not yet appropriate to do an impact evaluation (because one or more of the necessary conditions are not met), then channel your efforts and resources into evaluation studies that will tell you about the quality of the communication activities that make up the D&D component. Rather than focus on the changes brought about by the communication activities, examine their quality as judged by criteria such as conformance with principles of communication or regard for activities by their audiences. These kinds of information will allow you to make other kinds of judgments about the effectiveness of your demonstration and dissemination activities.

REFERENCES

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Cook, T. D., & Campbell, D. T. Quasi-experimentation: Design and analysis issues for field settings. Chicago: Rand McNally, 1979.

Glass, G. V., Willson, V. L., & Gottman, J. M. Design and analysis of time-series experiments. Boulder: University of Colorado Press, 1975.

Kerlinger, F. N. Foundations of behavioral research (2nd ed.). New York: Holt, 1973.

Stake, R. E. The countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

Tatsuoka, M. Multivariate analysis techniques for educational and psychological research. New York: Wiley, 1971.

Weiss, C. Evaluation research: Methods of assessing program effectiveness. Englewood Cliffs, NJ: Prentice-Hall, 1972.